CHAPTER 6

Probability Distributions

In Chapter 6, you will learn how to use R for several important probability distributions. We cover the binomial distribution, the Poisson distribution, the normal distribution, the *t* distribution, the F distribution, and the chi-square distribution. R uses a standardized method for naming the functions associated with each of these, so that it is easy to remember them. We use the letters *d* (for density), *p* (for probability), *q* (for quantile), and *r* (for random) to preface the function, so for the normal distribution, `dnorm` returns the density, `pnorm` returns the probability associated with an area under the curve, `qnorm` returns the score associated with a given probability, and `rnorm` produces the desired number of randomly generated scores from a normal distribution with a given mean and standard deviation.

Probability distributions may be *discrete*, as in the case of the binomial and Poisson distributions, or they may be *continuous*, as in the case of the normal, *t*, *F*, and chi-square distributions. We will consider discrete probability distributions first.

6.1 Discrete Probability Distributions

Probability can be defined in different ways. For example, we can define the probability of receiving a heads on the flip of a fair coin as 1/2 or .50. This is the theoretical probability, because the sample space has two possible outcomes, and there is only one outcome resulting in heads. We could also wait until we have tossed the coin a number of times to determine how many heads we got as a proportion of the total number of tosses. This is an empirical probability. Specifically, the probability of an event, *E*, is the relative proportion of the number of times *E* occurs as a proportion of the total number of observed occurrences.

There are any number of discrete probability distributions. A discrete random variable can take on only clearly separated values, such as heads or tails, or the number of spots on a six-sided die. The categories must be mutually exclusive and exhaustive. Every event belongs to one and only one category, and the sum of the probabilities is 1. We can in general calculate the mean and variance of a discrete probability distribution as follows, though as you will see, we can often simplify the calculations for certain distributions. First, the mean of any discrete probability distribution can be computed by the following:

$$\mu = \sum\left[xP(x)\right] \tag{6.1}$$

The variance for a discrete probability distribution is calculated as follows:

$$\sigma^2 = \sum\left[(x-\mu)^2 P(x)\right] \tag{6.2}$$

6.2 The Binomial Distribution

Statisticians rely on the law of large numbers, which tells us that when we perform the same experiment a large number of times, the average of the results obtained from repeated trials (the empirical value) should be close to the expected (theoretical) value. The law of large numbers was first derived by Swiss mathematician Jacob Bernoulli, for whom the Bernoulli process is also named. A Bernoulli process is a finite or infinite sequence of independent random variables X_1, X_2, X_3, . . . such that for each i, the value of X_i is either 0 or 1. An obvious, if perhaps unimaginative, form of a Bernoulli process is a coin-flipping experiment. The extension of the Bernoulli process to more than two possible outcomes is known as the Bernoulli scheme (e.g., the possible outcomes of throwing a six-sided die).

The discrete binomial distribution is very useful for modeling processes in which the binary outcome can be either a success (1) or a failure (0). The random variable X is the number of successes in N independent trials, for each of which the probability of success, p, is the same. The number of successes can range from 0 to N. The expected value of k is $N\,p$, the number of trials times the probability of success on a given trial, and the variance of the binomial distribution is $N\,pq$, where $q = 1 - p$. We calculate the binomial probability as follows:

$$p(X = k \mid p, N) = \binom{N}{k} p^k (1-p)^{N-k} \tag{6.3}$$

The binomial coefficient $\binom{N}{k}$ is not related to the fraction $\frac{N}{k}$, and it is often written ${}_nC_k$, read "N choose k." The binomial coefficient can be calculated as follows:

$$\binom{N}{k} = \frac{N!}{k!(N-k)!} \tag{6.4}$$

We can use the `choose()` function in R to find a binomial coefficient, for example, the number of ways you can select six individual objects from a total of 10. Simply type `choose(N, k)`, substituting the desired values:

```
> choose (10, 6)
[1] 210
```

We use the binomial distribution in many ways. In each case, the number of "successes" is counted, and we determine the probability of either an exact number of successes given n and p or the probability of a range of numbers of successes. As it materializes, the binomial distribution also gives us a good approximation of the normal distribution as the number of trials increases, and as the probability of success is close to .50. Here is a binomial distribution for the number of successes (heads) in 10 tosses of a fair coin, in which the probability of success for each independent trial is .50. We establish a vector of the number of successes (heads), which can range from 0 to 10, with 5 being the most likely value. The `dbinom` function produces a vector of values, but to make ours easier to read in a more customary tabular format, we use `cbind` again to create a matrix instead, changing the row names from the standard 1 to 11 to the more sensible 0 to 10 by using the `rownames` function.

```
> x <- 0:10
> x
[1]  0  1  2  3  4  5  6  7  8  9  10
> binomCoinToss <- cbind(dbinom(x,  10, .50))
> rownames(binomCoinToss) <- x
> binomCoinToss
          [ ,1]
0  0.0009765625
1  0.0097656250
2  0.0439453125
3  0.1171875000
4  0.2050781250
5  0.2460937500
6  0.2050781250
7  0.1171875000
8  0.0439453125
9  0.0097656250
10 0.0009765625

> class(binomCoinToss)
[1] "matrix"
```

Careful examination reveals that the distribution is symmetrical, and a plot of the distribution against the values of x makes this clearer. The shape of the distribution is close to "normal looking," even though the binomial distribution is discrete, as shown in the probability mass function (PMF) in Figure 6-1. Here's how I plotted the PMF and added the points and the horizontal line for the *x* axis. Somewhat confusingly, the type = "h" plots *vertical* lines on the graph. According to the R documentation, the h is for "histogram like" or "high-density" vertical lines.

```
> binomDist <- dbinom (x, 10, 0.50)
> plot (x, binomDist, type = "h")
> points (x, binomDist)
> abline (h = 0)
> lines (x, binomDist)
```

The addition of lines to "connect the dots" makes it more obvious that the distribution is "normal" in appearance. We find that the binomial distribution serves as a good approximation of the normal distribution as the probability of success gets closer to .50 and as the number of trials increases.

Traditionally, statistics books provided tables of various quantiles of the binomial probability distribution. Modern texts depend on some form of tech (such as the built-in functions in R) to render the tables unnecessary. For example, if we want to know the probability of any exact number of successes, we can use the `dbinom` function, and if we are interested in finding the probability of a range of successes, we can use the `pbinom` function. For example, what is the probability of throwing 5 or fewer heads in our 10 tosses of a fair coin? We could calculate the individual probabilities for 0, 1, 2, 3, 4, and 5 and then add them up, and that would produce the correct answer, but it would be wasteful, as we can use `pbinom` for the same answer as follows. Note that when we change the default `lower.tail = TRUE` argument to `lower.tail = FALSE`, we are getting a right-tailed probability. It is important to note that for the `pbinom` function, the boundary value is included in the lower-tailed probability interval but excluded from the upper-tailed interval, as the following code illustrates. Adding the last two probabilities should produce 1, as 5 cannot "be" in both intervals.

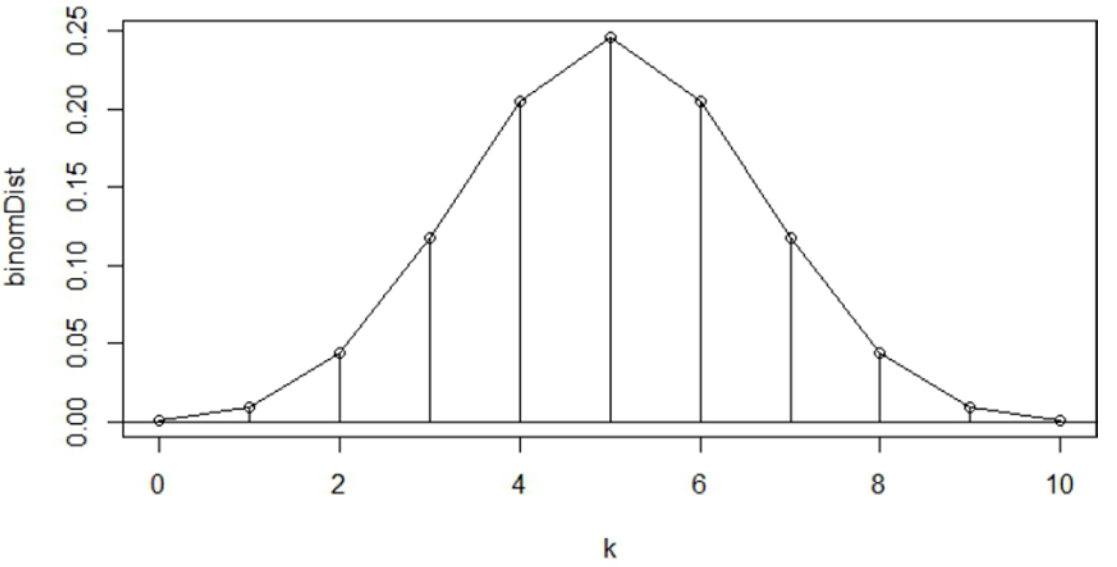

Figure 6-1. *Binomial distribution of the number of heads in 10 coin tosses*

```
> sum (dbinom(0:5, 10, .50))
[1] 0.6230469
> # calculate the probability k <= 5
> pbinom(5, 10, .50)
[1] 0.6230469
> # calculate the probability k > 5
> pbinom(5, 10, .50, lower.tail = FALSE)
[1] 0.3769531
```

The problem of whether to include the endpoint or not is important only with discrete probability distributions, because with continuous distributions, the probability of an exact point on the probability density curve is essentially zero.

Let us perform a simulation using the `rbinom` function to see how it works. We will also get an idea of whether the law of large numbers mentioned earlier can be demonstrated effectively. We will simulate the tossing of a fair coin 1,000 times, recording whether the coin is heads (1) or tails (0) for each coin toss. We will also plot the running average of the number of heads against the trial number. We can use the `cumsum()` function to keep track of the number of heads, as heads are 1s and tails are 0s, and then we can calculate the running averages as follows. We can then plot the proportion of heads over the series of trials and add a horizontal reference line at $p = .50$ (see Figure 6-2)

```
> tosses <- rbinom(1000, 1, .5)
> heads <- cumsum(tosses)/1:1000
> plot(heads, type = "l", main = "Proportion of Heads")
> abline (h = .5)
```

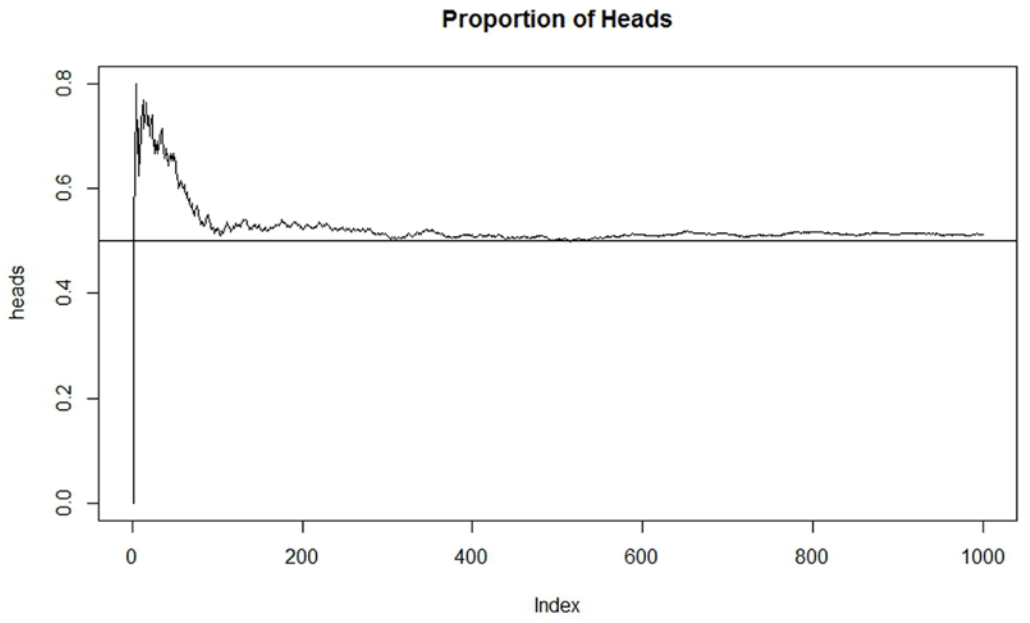

Figure 6-2. *Proportion of heads in 1,000 simulated coin tosses*

6.2.1 The Poisson Distribution

The Poisson distribution is a special case of the binomial distribution. We define success and failure in the usual way as 1 and 0, respectively, and as with the binomial distribution, the distinction is often arbitrary. For example, it makes little sense to talk about a hurricane or a work-related death as a "success." For that reason, we will change the word "success" to occurrence. The Poisson distribution, unlike the binomial distribution, has no theoretical upper bound on the number of occurrences that can happen within a given interval. We assume the number of occurrences in each interval is independent of the number of occurrences in any other interval. We also assume the probability that an occurrence will happen is the same for every interval. As the interval size decreases, we assume the probability of an occurrence in the interval becomes smaller. In the Poisson distribution, the count of the number of occurrences, X, can take on whole numbers 0, 1, 2, 3, … The mean number of successes per unit of measure is the value μ. If k is any whole number 0 or greater, then

$$P(X = k) = \frac{e^{-\mu}\mu^{k}}{k!} \tag{6.5}$$

The variance of the Poisson distribution is also μ, and the standard deviation is therefore $\sqrt{\mu}$. As with the binomial distribution, we have the `dpois`, the `ppois`, the `qpois`, and the `rpois` functions.

As an example, the U.S. Bureau of Labor Statistics reported that in 2012, a year for which complete data are available, the number of work-related deaths per day averaged 12.7. Assuming the count of work-related deaths follows a Poisson distribution, what is the probability of exactly 10 deaths on a given day? What is the probability of 10 or fewer deaths in one day? What is the probability of more than 10 deaths in a day? Just as with the binomial distribution, the discrete Poisson distribution includes the boundary value in the lower-tailed probability and excludes it from the upper-tailed probability.

```
> dpois(10, 12.7)
[1] 0.09177708
> ppois(10, 12.7)
[1] 0.2783314
> ppois(10, 12.7, lower.tail = FALSE)
[1] 0.7216686
```

6.2.2 Some Other Discrete Distributions

In addition to the binomial and the Poisson distributions, there are other useful discrete probability distributions. The negative binomial distribution is one of these. Instead of determining the distribution of successes over a fixed number of trials, as we do in the binomial distribution, we determine the number of failures that are likely to occur before a target number of successes occur. The negative binomial distribution is built into R as well. A special case of the negative binomial distribution is the geometric distribution, which is the distribution of the number of failures that occur before the first success. As you would suspect, this distribution is built into R, too. The nice thing about the discrete probability distributions as well as the continuous probability distributions in R is that in a sense once you have learned one, you have learned them all, as the standardized function naming makes it easy to understand how to look up probabilities, how to find areas, and how to do reverse-lookups, that is, how to find the value associated with a given probability.

6.3 Continuous Probability Distributions

Continuous variables can take on any value within some specified range. Thus continuous probability functions plot a probability density function (PDF) instead of a discrete probability mass function (PMF). In contrast to discrete probability distributions, the probability of a single point on the curve is essentially zero, and we rarely examine such probabilities, rather focusing on areas under the curve. In statistics, the four most commonly used continuous probability distributions are the normal distribution and three other distributions theoretically related to the normal distribution, namely, the *t* distribution, the *F* distribution, and the chi-square distribution.

6.3.1 The Normal Distribution

The normal distribution serves as the backbone of modern statistics. As the distribution is continuous, we are usually interested in finding areas under the normal curve. In particular, we are often interested in left-tailed probabilities, right-tailed probabilities, and the area between two given scores on the normal distribution. There are any number of normal distributions, each for any non-zero value of σ, the population standard deviation, so we often find it convenient to work with the unit or standard normal distribution. The unit normal distribution has a mean of 0 (not to be confused in any way with a zero indicating the absence of a quantity), and a standard deviation of 1. The normal distribution is symmetrical and mound shaped, and its mean, mode, and median are all equal to 0. For any normal distribution, we can convert the distribution to the standard normal distribution as follows:

$$z = \frac{(x - \mu_x)}{\sigma_x} \tag{6.6}$$

which is often called z-scoring or standardizing. The empirical rule tells us that for mound-shaped symmetrical distributions like the standard normal distribution, about 68% of the observations will lie between plus and minus 1 standard deviation from the mean. Approximately 95% of the observations will lie

within plus or minus 2 standard deviations, and about 99.7% of observations will lie within plus or minus 3 standard deviations. We can use the built-in functions for the normal distribution to see how accurately this empirical rule describes the normal distribution. We find the rule is quite accurate.

```
> pnorm(3) - pnorm(-3)
[1] 0.9973002
> pnorm(2) - pnorm(-2)
[1] 0.9544997
> pnorm(1) - pnorm(-1)
[1] 0.6826895
```

By subtracting the area to the left of the lower *z* score from the area to the left of the higher *z* score, we retain the area between the two scores. The `qnorm` function can be used to locate precise *z* scores that correspond to a given probability. For example, in statistics we commonly accept .05 or .01 as our standard for statistical significance. To find the critical values of *z* that will put half of the alpha level in the upper tail and half in the lower tail, we find the *z* score associated with a probability of $1 - \alpha/2$. Let's start with an alpha level of .05, and then find the critical values for .01 and .10 as well:

```
> qnorm(1 - .05/2)
[1] 1.959964
> qnorm(1 - .01/2)
[1] 2.575829
> qnorm(1 - .10/2)
[1] 1.644854
```

Of course, it is more direct just to type `qnorm(.975)`, but the listing makes it obvious to the reader what we are really doing. These are standard critical values of *z* that can be found in any table of the standard normal distribution, but the advantage of R is that it makes such tables unnecessary, as we can find critical values more quickly and more accurately with technology than by reading printed tables. To prove the point, let's determine the area between *z* scores of +1.96 and -1.96 using our previous subtraction strategy. Indeed the area is almost exactly .95 or 95%.

```
> pnorm(1.96) - pnorm(-1.96)
[1] 0.9500042
```

If we are doing a one-tailed test, we place the entire alpha level in one tail of the standard normal distribution. Conducting a one-tailed test has the simultaneous effect of making the statistical test more powerful given that the results are in the hypothesized direction and making it technically inappropriate to talk about findings that are not in the hypothesized direction. The default for most statistical software is to perform a two-tailed test, but it is possible also to specify a left-tailed or a right-tailed test as well.

One of the most important applications of the normal distribution is its ability to describe the distributions of the means of samples from a population. The central limit theorem tells us that as the sample size increases, the distribution of sample means becomes more and more normal, regardless of the shape of the parent distribution. This is the statistical justification for using the normal distribution and theoretically related distributions such as the *t*, *F*, and chi-square distribution, for tests on means, proportions, and deviations from expectation.

The `faithful` dataset supplied in the base version of R shows the distributions of the duration of the eruptions of the Old Faithful geyser and the waiting times between eruptions (Figure 6-3). Both measurements are in minutes.

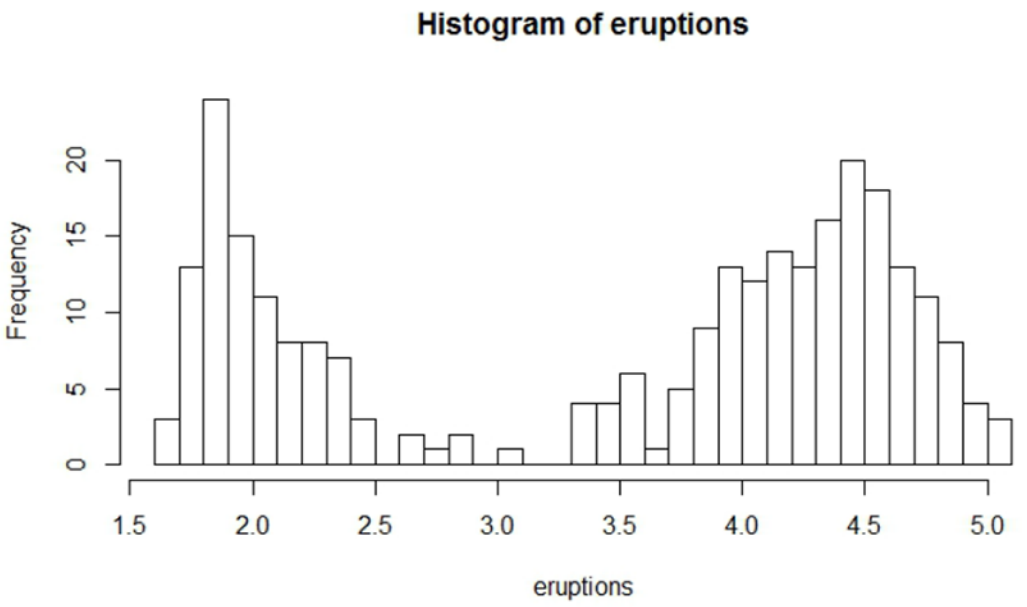

Figure 6-3. *Duration of eruptions of Old Faithful geyser*

Let us illustrate the central limit theorem (CLT) using this unusual dataset. We will first take samples of size 5 and calculate their means, and then we'll repeat the process with samples of size 20 for the sake of comparison. If the CLT is correct, the shape of the sampling distribution of means should approach a normal distribution as the sample size increases, even though the parent distribution is far from normal.

We take $N = 999$ samples of size 5 from the eruption data, calculating the mean for each sample, and saving the means using the `replicate()` function. The `replicate()` function runs R code (an expression) N times and is as if we typed and re-ran the code manually N times. We can then repeat the process with a sample size of 20. Using the `par()` function allows us to control the graphics output so that we can show the histograms with the sampling distributions of means for the two sample sizes in side-by-side comparison (see Figure 6-4).

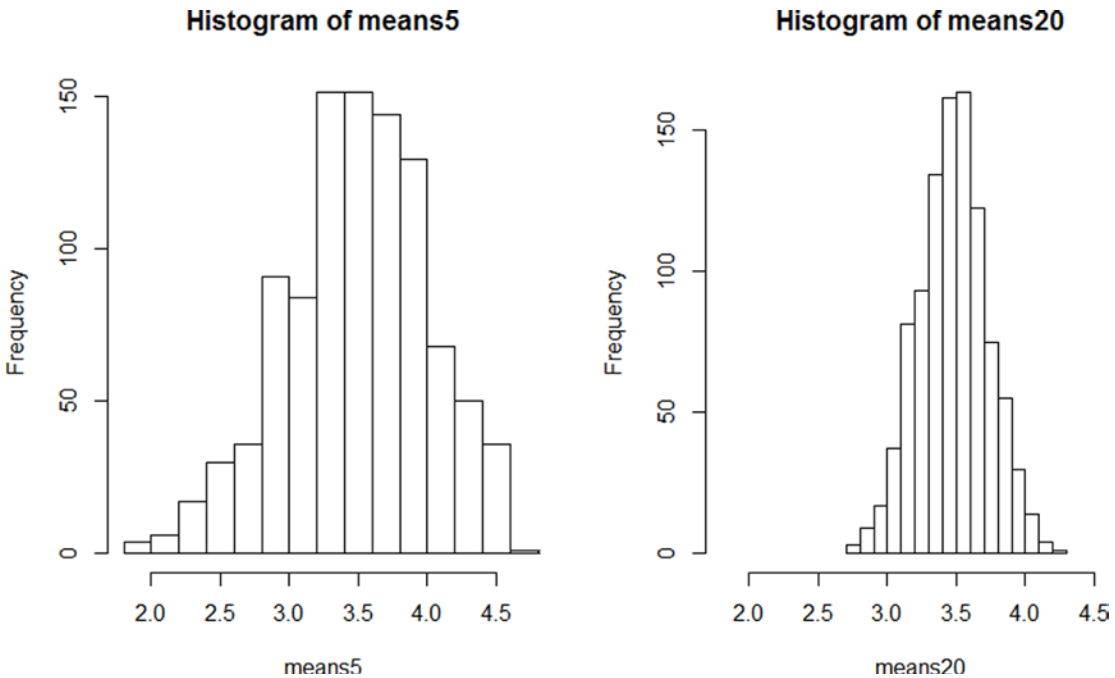

Figure 6-4. *Sample distributions of means for samples of size 5 and 20*

Observation of the histograms reveals two important things. First, the spread of the sample means is larger with small sample sizes. Second, as promised by the CLT, we see that as the sample size increases, the shape of the sampling distribution becomes more normal looking. This gives us the reassurance we need to rely on the normal distribution as a descriptor of the shape of sampling distributions of means, even from uniform distributions, skewed distributions, or even bimodal distributions like that of the eruption data.

```
> attach(faithful)

> sampsize5 <- 5
> means5 <- replicate(999, mean(sample(eruptions, sampsize5, replace = TRUE)))

> sampsize20 <- 20
> means20 <- replicate(999, mean(sample(eruptions, sampsize20, replace = TRUE)))
> par (mfrow=c(1,2))

> hist (means5, breaks = 15, xlim = c(1.8, 4.7))
> hist (means20, breaks = 15, xlim = c(1.8, 4.7))

> detach(faithful)
```

6.3.2 The *t* Distribution

Although the mathematical function for the PDF of the *t* distribution is quite different from that for the normal distribution, the *t* distribution approaches the normal distribution as the degrees of freedom increase. The degrees of freedom parameter is based on the sample size. The *t* distribution was developed as a way to examine the sampling distribution of means for small samples, and it works more effectively for that purpose than does the normal distribution.

Most statistical software programs, including R and SPSS, use the *t* distribution for calculating confidence intervals for means. As the examination of the listing reveals, the *t* distribution is definitely needed for small sample sizes, and it works adequately for large sample sizes. The normal distribution, on the other hand, is appropriate only for situations in which the population standard deviation is known or where the sample size is large.

If we were to develop a 95% confidence interval for the mean using the standard normal distribution, we would find that the critical values of ±1.96 would apply in all cases. With the *t* distribution, the critical values would vary with the degrees of freedom. Critical values of *t* for various one- and two-tailed hypothesis tests were once located in the backs of most statistics texts, but as with the other probability distributions, the tables are not necessary when one has access to R. The built-in functions for the *t* distribution work in the same way as those for the other probability distributions. We can determine the exact two-tailed probability or one-tailed probability for a given value of *t*, or the critical value for any one- or two-tailed hypothesis test. The critical values of *t* for a 95% confidence interval with 18 degrees of freedom are found as follows:

```
> qt(0.975, 18)
[1] 2.100922
```

To find a critical value for *t*, we must use the same strategy we use for the normal distribution. However, we must supply the degrees of freedom parameter as well. As with the normal distribution, we use the strategy of placing half of alpha in each tail of the *t* distribution for a two-tailed test or a confidence interval. With one-tailed tests, we place all of alpha in the upper or lower tail. For example, with a right-tailed test and 18 degrees of freedom, we find the critical value placing all of alpha in the right tail. Because the *y* axis never touches the *x* axis, there are no theoretical upper or lower bounds to the *t* or *z* distributions. In most programs or calculators one can substitute an arbitrarily large number such as ±999 for the lower or upper bound, as the *z* scores or *t* values at such extremes are essentially zero. R, however, provides the `Inf` and `-Inf` objects to represent positive and negative infinity. We see that indeed the value of 1.734 cuts off the upper 5% of the *t* distribution for 18 degrees of freedom from the lower 95%.

```
> qt (0.95, 18)
[1] 1.734064
> pt (1.734064, 18) - pt(-Inf, 18)
[1] 0.95
```

As mentioned earlier, the *t* distribution converges on the standard normal distribution as the degrees of freedom become larger, making the differences between the two smaller and smaller. Let us plot the standard normal distribution and then overlay the *t* distributions for 1, 4, and 9 degrees of freedom. We see that even with a sample size of 10, the *t* distribution becomes "normal looking" very quickly (see Figure 6-5).

```
xaxis <- seq(-4, 4, .05)
y <- dnorm(xaxis)
y1 <- dt(xaxis, 1)
y4 <- dt(xaxis, 4)
y9 <- dt(xaxis, 9)
plot(xaxis, y, type = "l")
lines(xaxis, y1, col = "purple")
lines(xaxis, y4, col = "red")
lines(xaxis, y9, col = "blue")
```

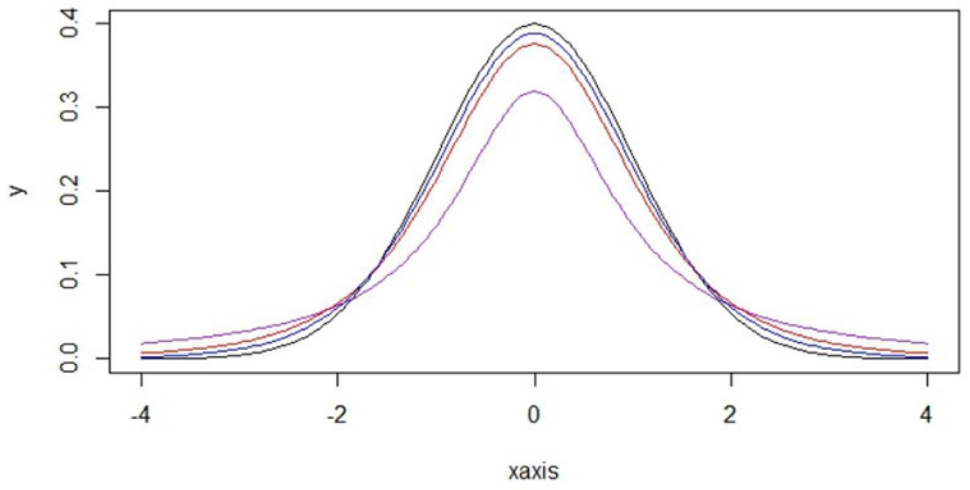

Figure 6-5. *Comparison of the standard normal and t distributions*

6.3.3 The *F* distribution

The *F* distribution, like the *t* distribution, has degrees of freedom, but in this case, because *F* is the ratio of two variances or variance estimates, there are degrees of freedom for both the numerator term and the denominator term. Again, traditionally, statistics textbooks included tables of critical values of *F* for varying combinations of degrees of freedom, but as before, these are found to be unnecessary when R or other technology is available and those tables are more and more likely to be found absent.

The *F* distribution is positively skewed, and for most purposes, we place the critical values in the upper tail by following the expedient of dividing the larger variance estimate by the smaller variance estimate, though it is entirely possible to have a left-tailed critical value of *F*. As the degrees of freedom increase, the *F* distribution become asymptotically normal. Let's produce *F* distributions for several combinations of degrees of freedom, using the `ylim` argument to specify the limits of the *y* axis:

```
> xaxis <- seq(0, 8, .05)
> y1 <- df(xaxis, 3, 5)
> y2 <- df(xaxis, 6, 10)
> y3 <- df(xaxis, 9, 20)
> y4 <- df(xaxis, 49, 49)
> plot(xaxis, y1, type = "l", xlab = "Value of F", main = "PDF of F Dist.",
  ylim = c(0, 1.5), col = "green")
> lines (xaxis, y2, col = "red")
> lines (xaxis, y3, col = "blue")
> lines (xaxis, y4, col = "purple")
```

The plot shows that as the degrees of freedom increase, the *F* distribution becomes more symmetrical and clusters around a value of 1 (see Figure 6-6). This is because when the null hypothesis is true that the variances being compared are equal, the value of the *F* ratio would be 1.

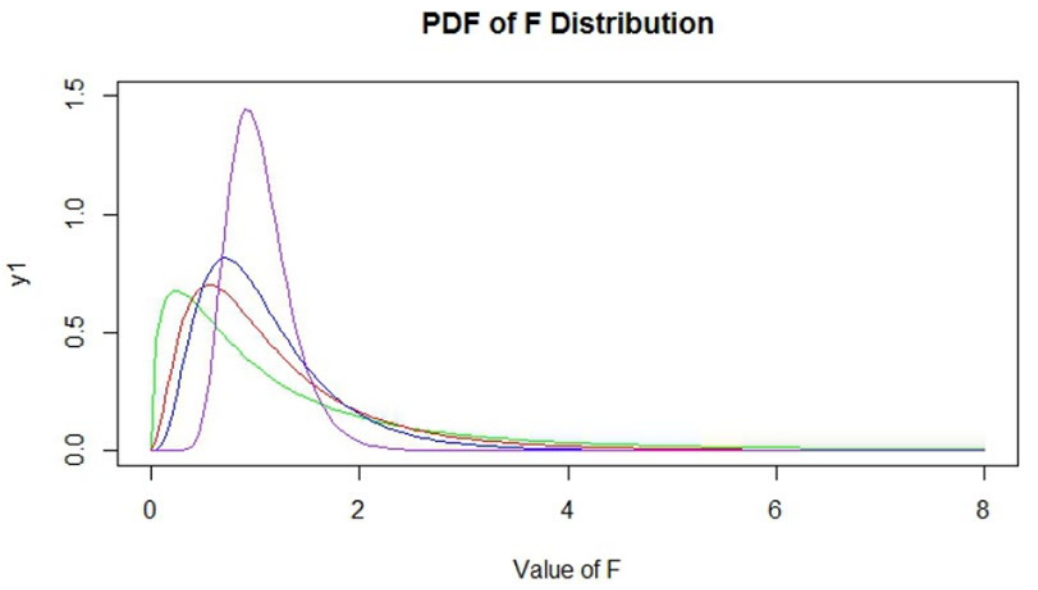

Figure 6-6. *The F distribution becomes more symmetrical as the degrees of freedom increase*

6.3.4 The Chi-Square Distribution

The chi-square distribution has one parameter, the degrees of freedom. We use the chi-square distribution for tests of frequency counts and cross- tabulations. The chi-square distribution is also very useful for tests involving model fit. Like the *F* distribution, the chi-square distribution is positively skewed. Figure 6-7 shows the chi-square distributions for varying degrees of freedom. We will discuss it in more detail in our chapters on graphics (Chapter 9) and data visualization (Chapter 17), but here I introduce the addition of text to a plot by use of the `text` function. I used the built-in `locator()` function to determine the (*x, y*) coordinates where I wanted the labels to be placed. The default is to center the label at the coordinate pair, but the `adj = c(0,0)` argument begins the label at the coordinate pair.

```
> xaxis <- seq(0, 20, .05)
> y1 <- dchisq(xaxis, 4)
> y2 <- dchisq(xaxis, 6)
> y3 <- dchisq(xaxis, 10)
> plot(xaxis, y1, type = "l", xlab = "Chi - square Value")
> lines(xaxis, y2, col = "blue")
> lines(xaxis, y3, col = "red")
> xcoords <- c(3.4, 5.75, 10.6)
> ycoords <- c(0.17, 0.13, 0.09)
> labels <- c("df = 4", "df = 6", "df = 10")
> text(xcoords, ycoords, labels, adj = c(0,0))
```

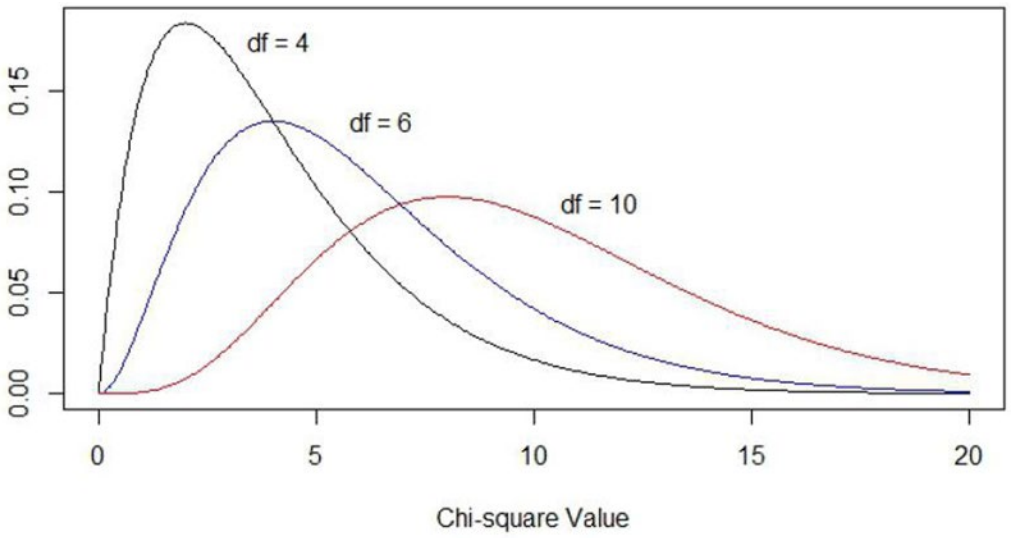

Figure 6-7. *The chi-square distribution also becomes more symmetrical as the degrees of freedom increase*

As with the *F* distribution, the chi-square distribution shifts rightward and becomes more mound shaped as the degrees of freedom increase. The mean of any chi-square distribution is equal to its degrees of freedom, and the mode is equal to the degrees of freedom minus 2.

The chi-square distribution is usually attributed to Karl Pearson in his development of tests of goodness of fit. In truth, Pearson independently rediscovered a distribution identified by German statistician Friedrich Robert Helmert a quarter-century earlier. Helmert described the chi-square distribution in relation to the distribution of the sample variance. Thus, the chi-square distribution can also be used in hypothesis tests and confidence intervals for the variance and standard deviation. With the assumption that a sample is drawn from a normal population, we can test the hypothesis that the population variance is equal to some specified value, σ^2, by calculating the following chi-square statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{6.7}$$

With a firm grasp on the most common discrete and continuous probability distributions, we are now ready to discuss the analysis of tables, including chi-square tests of both goodness of fit and independence, in Chapter 7.

■ **Note** In this chapter, we mentioned critical value(s) which may be equivalently used in null hypothesis testing in place of *p* value. Most modern introductory statistical textbooks have deemphasized critical value in favor of *p* value. There are even modern practitioners who choose not to use either as all-or-nothing rules to accept or reject the null hypothesis. In this text, we will limit ourselves to simply describing both critical and *p*-value R coding.

References

1. J. T. Roscoe, *Fundamental Research Statistics for the Behavioural Sciences*, 2nd ed. (New York: Holt, Rinehart & Winston, 1975).