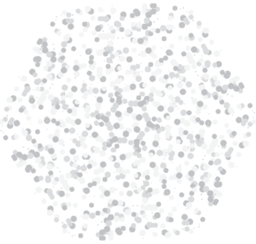
9



Hypothesis testing with the CLT

> I'm looking for the patterns in static: They start to make sense the longer I'm at it.
>
> —Gibbard (2003)

The purpose of descriptive statistics is to say something about the data you have. The purpose of hypothesis testing is to say something about the data you don't have.

Say that you took a few samples from a population, maybe the height of several individuals, and the mean of your sample measurements is $\hat{\mu} = 175$ cm. If you did your sums right, then this is an indisputable, certain fact. But what is the mean height of the population from which you drew your data set? To guess at the answer to this question, you need to make some assumptions about how your data set relates to the population from which it was drawn.

Statisticians have followed a number of threads to say something about data they don't have. Each starts with a data set and some assumptions about the environment and data generation method, and concludes with an output distribution that can be compared to the data. Here is a list of some common assumptions. It is impossible for it to be comprehensive, and many of the categories overlap, but it offers a reasonable lay of the discussion in the following chapters.

- *Classical methods*: Claim that the data was produced via a process that allows application of the Central Limit Theorem.

- *Maximum likelihood estimation*: Write down a likelihood function for any given data/parameter pairing, and find the most likely parameter given the data.
- *Bayesian analysis*: Claim a distribution expressing prior beliefs about the parameters and a likelihood function for the data on hand, then combine them to produce a posterior distribution for the parameters.
- *Resampling methods*: Claim that random draws from the data are comparable to random draws from the population (the *bootstrap principle*), then generate a distribution via random draws from the data.
- *Kernel/smoothing methods*: Claim that the histogram of the existing data is a lumpy version of the true distribution; smooth the data to produce the output distribution.

All of these approaches will be discussed over the remainder of this book. This chapter will focus on the first: making inferences about the population via use of the Central Limit Theorem (*CLT*). The CLT describes the distribution of the sample mean, $\overline{\mathbf{x}}$, and works regardless of the form of the underlying data. That is, no matter the true distribution of the data, the distribution of the sample mean has a very specific form—as long as n is large enough. For relatively small n, another of the above methods of inference, such as the Monte Carlo methods discussed in Chapter 11, may be preferable.

Metadata

Metadata is data about data. Any statistic is a function of data, so it is by definition metadata. Be careful not to confuse the characteristics of the data and metadata; for example, the variance of the mean is almost always smaller than the variance of the base data. Like many hypothesis tests, the Central Limit Theorem is primarily concerned not with the distribution of the base data set, but the distribution of the mean of the data set.

The CLT gives us a basis for the Normal distribution; we can then produce variants based on the Normal. The square of a Normally distributed variable x will have a Chi squared distribution (which is written as $x^2 \sim \chi^2_1$, and read as: the statistic is distributed as a Chi squared with one degree of freedom). Dividing a Normal distribution by a transformed χ^2 distribution produces another distribution (the t distribution), and the ratio of two χ^2's produces an F distribution. Because all of this is rooted in the CLT, the statements are true regardless of the vagaries of the underlying population that the statistics describe.

Having found a means of describing the distribution of the unobservable population parameter β, Section 9.3 will then look a number of simple tests regarding β. They are direct applications of the above distributions, and so are often given names like the t test, χ^2 test, and F test.

The remainder of the chapter applies these building blocks in more complex structures to test hypotheses about more elaborate statistics. For example, if two independent statistics β_1 and β_2 are $\sim \chi^2_1$, then $\beta_1 + \beta_2 \sim \chi^2_2$. So if the squared distance between a histogram segment and a hypothesized distribution is $\sim \chi^2_1$, then the total distance between a thousand such segments and the hypothesized distribution is $\sim \chi^2_{1000}$, and that total distance could thus be used to test the aggregate claim that the data histogram is close to the distribution.

9.1 THE CENTRAL LIMIT THEOREM

The CLT is the key piece of magic for this chapter. Make a series of n *independent*, *identically distributed* draws, $x_1, x_2, \dots x_n$, from a fixed underlying population. The underlying population may have *any* nondegenerate distribution.[1] Let the mean of this sequence of draws be $\overline{\mathbf{x}}$, and the true mean of the overall population be μ. Then as $n \to \infty$,

$$\sqrt{n}\frac{(\overline{\mathbf{x}} - \mu)}{\sigma} \sim \mathcal{N}(0, 1). \tag{9.1.1}$$

That is, no matter the underlying population, the distribution of a mean of a series of draws will approach a Normal distribution.

Put another way, if we repeated the procedure and drew k independent data sets from the population and plotted a histogram of $\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_k$, we would eventually see a familiar bell curve.

Because it is about the distribution of $\overline{\mathbf{x}}$, the CLT embodies a two-stage procedure: we first produce the means of a series of k sets of draws—metadata from the base distribution—and then seek the distribution of the metadata (the k means), not the data itself. Listing 9.1 demonstrates exactly how this two-stage procedure works, and is worth understanding in detail.

- On line four (and the first panel of Figure 9.2), you can see the data from which the program makes draws. It is nowhere near a bell curve: everything is either ≤ 11 or ≥ 90.
- The inner loop of the `make_draws` function (the `j`-indexed loop) takes `ct` draws from the CDF, and adds them to `total`. When `total` is divided by `ct` in the line after the loop, it becomes the mean of `ct` draws from the distribution. The outer loop (the `i`-indexed loop) records `drawct` such means. Line 23 plots the distribution of this set of several means.
This double-loop is the base of the CLT, and is reflected in the assumptions about data sets below. Say that we have `drawct` data points in our data set, and we are

[1] By 'nondegenerate' I mean that more than one outcome has positive probability. If your data is fifty items that all have value 10, the mean of your samples will be 10 no matter how you slice them. However, if your sample takes as few as two distinct values, the CLT will eventually provide you with a bell curve. You can verify this by modifying the code in Listing 9.1. There are also theoretical distributions with infinite variance, which also cause problems for the CLT, but this is of course not an issue for finite data sets.

```
 1 #include <apop.h>
 2
 3 int drawct = 10000;
 4 double data[] = {1, 2, 3, 10, 11, 10, 11, 90, 91, 90, 91};
 5
 6 gsl_vector *make_draws(int ct, gsl_rng *r){
 7   double total;
 8   gsl_vector *out = gsl_vector_alloc(drawct);
 9     for(int i=0; i< drawct; i++){
10         total = 0;
11         for(int j=0; j< ct; j++)
12             total += data[gsl_rng_uniform_int(r, sizeof(data)/sizeof(data[0]))];
13         gsl_vector_set(out, i, total/ct);
14     }
15     return out;
16 }
17
18 int main(){
19   gsl_rng *r = apop_rng_alloc(23);
20     for (int ct=1; ct<= 1018; ct+=3){
21         printf("set title 'Mean of %i draws'\n", ct);
22         gsl_vector *o =make_draws(ct, r);
23         apop_plot_histogram(o, 200, NULL);
24         gsl_vector_free(o);
25         printf("pause 0.6\n");
26     }
27 }
```

Listing 9.1 Take the mean of an increasing number of draws. The distribution of the means approaches a Normal distribution. Online source: `cltdemo.c`.

claiming that they are Normally distributed. We presume that they are Normally distributed (and not just constant) because a multitude of events have affected each data point in some sort of haphazard way. That is, each individual data point went through a process like the inner loop in lines 11–12, absorbing a large number of random shocks. After all the little shocks, we gathered a single data point, as in line 13.

- The `main` function is intended to show that the CLT works best when each data point is the mean of several draws from the base distribution. Line 22 repeatedly calls `make_draws`. At the first call, `ct==1`, so `make_draws` makes 10,000 draws from the distribution itself. The next call produces 10,000 data points where each is the mean of four draws, and so on up to each data point being the mean of 1018 draws. The program dumps plots of the histograms to `STDOUT`, so run the program via `./cltdemo | gnuplot`.

Figure 9.2 shows a few frames of output from the program. The first frame of the animation is simply a set of spikes representing the base data. The second frame, where each data point is the mean of four draws, has a series of humps, because some draws have all large numbers, some have three large numbers and one small, some have two of each, and so on. In the third frame, there are more combinations possible, so there are more humps. As the frames progress, the humps merge together to form a familiar bell curve. This is a re-telling of the counting story on page 237, which explained why the Binomial distribution approaches a bell curve as well.

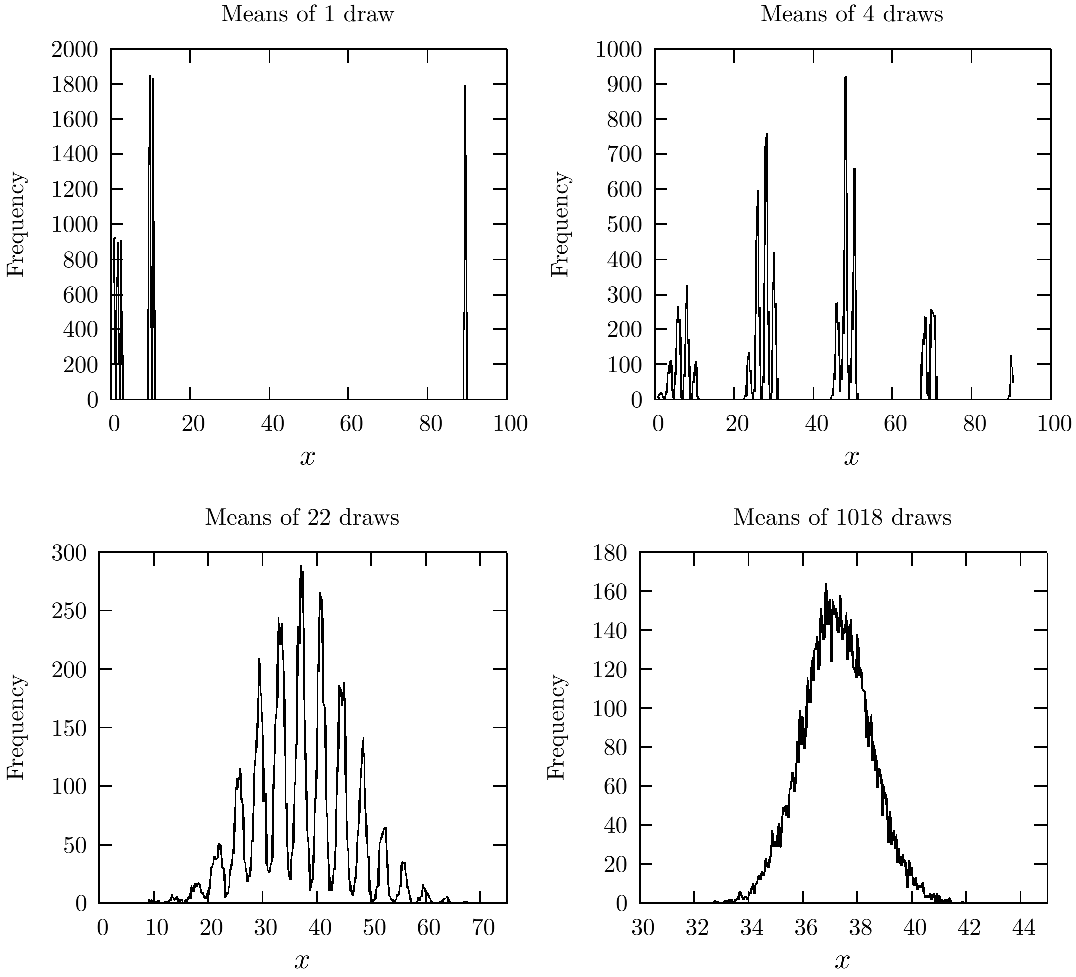


Figure 9.2 Sample outputs from the CLT demo in Listing 9.1

Finally, notice the x-axes of the snapshots: the original data was plotted from 0–100, but the scale in the fourth frame only goes from 30 to 45.[2] So not only does the distribution of $\overline{\mathbf{x}}$ approach a bell curve, but it approaches a rather narrow bell curve.

$\mathbb{Q}_{9.1}$

> Modify line 4 to try different base distributions from which the system will draw. [Thanks to the creative use of `sizeof` on line 12, you don't need to specify the size of the array. But see the footnote on page 125 on why this could be bad form.] Deleting the elements $\{1, 2, 3\}$ produces some especially interesting patterns. What sort of data sets lead quickly to a bell curve, and what data sets require averaging together thousands of elements before achieving a decent approximation to the Normal?

Equation 9.1.1 put the story more formally: if we have a data set $\mathbf{x}$ with n elements, true mean μ, and variance σ^2, then as $n \to \infty$, $(\bar{\mathbf{x}} - \mu) \big/ \frac{\sigma}{\sqrt{n}}$ approaches a $\mathcal{N}(0, 1)$ distribution. From that regularity of nature, we can derive all of the distributions to follow.

Variants on variance

There is often confusion about when to use σ^2, σ, or $\sigma/\sqrt{n}$, so it is worth a quick review.

- For any data set or distribution, the variance is notated as σ^2. The formula for the variance of a data set is $\sum_{i=1}^{n}(x_i - \mu)^2/n$, so it makes sense that its symbol would have a square included.
- For the Normal distribution, the square root of the variance is known as the standard deviation, σ, and is used to describe the 'width' of the distribution. For example, just over 95% of a Normal distribution is within plus or minus two standard deviations of the mean. Outside of the Normal distribution, σ is rarely used.
- Let us say that we have a *data set* $\mathbf{x}$ whose variance is σ^2. Then the variance of the *mean of* $\mathbf{x}$, $\mathrm{var}(\overline{\mathbf{x}})$, is σ^2/n, and the standard deviation of $\overline{\mathbf{x}}$ is $\sigma/\sqrt{n}$. Once again, it is important to bear in mind whether you are dealing with data or metadata.

Σ

> ➤ The Central Limit Theorem states that if each observation $\overline{\mathbf{x}}_i$ is the mean of some draws from an iid distribution, then as $n \to \infty$, the distribution of $\overline{\mathbf{x}}_i$ follows Equation 9.1.1.
>
> ➤ That is, if μ is the overall mean and σ is the square root of the variance of the set of $\overline{\mathbf{x}}_i$'s, then $(\overline{\mathbf{x}}_i - \mu)/\left(\sigma/\sqrt{n}\right)$ approaches a $\mathcal{N}(0, 1)$ distribution.

[2] If the shift in x-axis bothers you, you could ask Gnuplot to hold a constant scale by adding a line like `printf("set xrange [0:100]\n");` at the top of `main`.

9.2 Meet the Gaussian family

With the exception of the Normal, the distributions below are distinct from the distributions of Section 7.2. The distributions there are typically used to describe data that we have observed in the real world. The distributions here are aimed at describing metadata, such as the means and variances of model parameters.

Normal The Normal distribution, presented on page 241, will also be used to describe some of the parameters derived below. The big problem with the Normal distribution is that it depends on σ, an unknown. It also depends on μ, but we are frequently testing a claim that μ has some fixed value, so we assume rather than derive it. Thus, much of the trickery in this section involves combining distributions in ways such that the unknown σ's cancel out.

χ^2 distribution The square of a variable with distribution $\mathcal{N}(0,1)$ has a χ^2 distribution with one degree of freedom, and the sum of n independent χ^2-distributed variables is also $\sim \chi^2$, with n degrees of freedom. Figure 9.3 shows the distribution for a few different degrees of freedom.

- If $X \sim \mathcal{N}(0,1)$, then $X^2 \sim \chi_1^2$.
- If $X_i \sim \mathcal{N}(0,1)$ for $i = 1, \dots, n$, then $X_1^2 + \cdots + X_n^2 \sim \chi_n^2$.
- If $X_i \sim \chi_n^2$, then $E(X_i) = n$.

The summation property is immensely useful, because we often have sums of variables to contend with. The most common case is the sample variance, which is a sum of squares. Being a sum of squares of Normally-distributed elements, it is easy to show that (Snedecor & Cochran, 1976, p 74)

$$\frac{\left[\sum_n (x - \overline{x})^2\right]}{\sigma^2} \sim \chi_{n-1}^2. \tag{9.2.1}$$

The numerator is the estimate of the sample variance times $n-1$, so we can use this to test that the sample variance equals a given σ^2, or establish a confidence interval around an estimate of the variance. But we will see that it is useful for much more than just describing variance estimates.

The sample variance is $\sim \chi_{n-1}^2$, not χ_n^2, because given the first $n-1$ data points and the mean, the last one can actually be calculated from that data, meaning that we effectively have the sum of $n-1$ variables $\sim \chi_1^2$, plus a no longer stochastic constant. For more on degrees of freedom, see the sidebar on page 222.

It is worth mentioning the origin of the χ^2 distribution as a common form. Pearson (1900) did a Taylor expansion of errors from what we now call a Multinomial

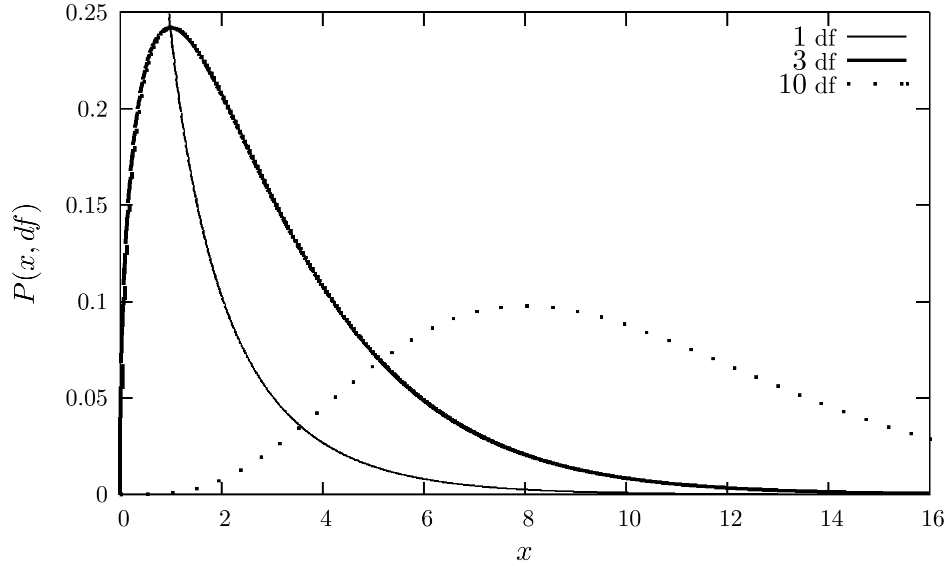


Figure 9.3 The χ^2 distribution for 3, 5, and 10 degrees of freedom. Being a sum of squares, it is always greater than zero. As more elements are added in to the sum, the distribution becomes less lopsided, and approaches a Normal distribution.

distribution, of the form $k_1\chi + k_2\chi^2 + k_3\chi^3 + \cdots$, where the k_i's are appropriate constants. He found that one can get satisfactory precision using only the χ^2 term of the series. The ANOVA family of tests is based on this approximation, because those tests claim that the data are random draws that fit the story of the Multinomial distribution (as on page 240), so a sum of such distributions leads to a χ^2 distribution as well.

STUDENT'S t DISTRIBUTION Let $\mathbf{x}$ be a vector of data (such as the error terms in a regression). Then

$$\frac{\overline{\mathbf{x}} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1},$$

where $\hat{\sigma} = \mathbf{x}'\mathbf{x}/n$ (a scalar). It might look as though this is just an approximation of the Normal, with $\hat{\sigma}$ replacing σ, but it is not. To see where the form of the t distribution came from, consider dividing the CLT equation (Equation 9.1.1),

$$\frac{\overline{\mathbf{x}} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1),$$

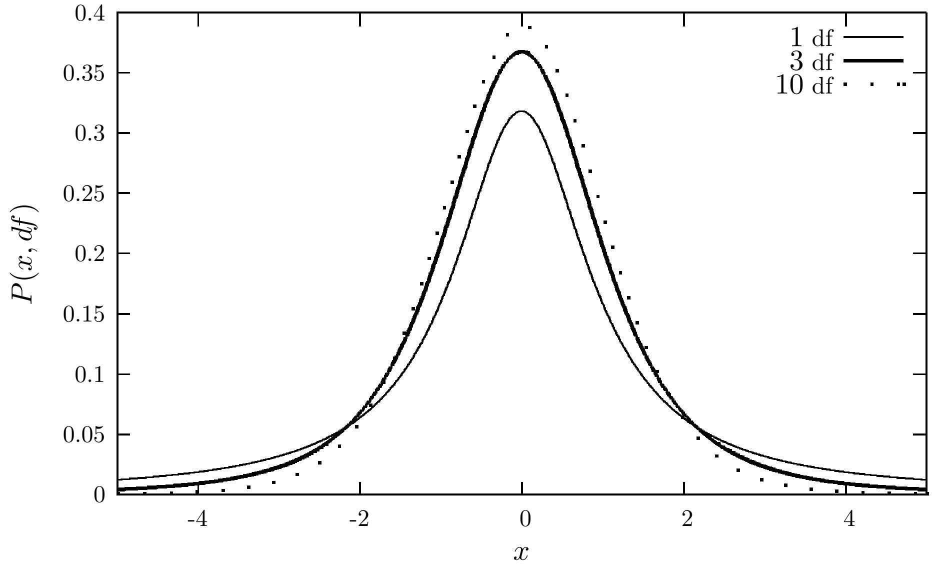


Figure 9.4 The t distribution for 1, 3, and 10 degrees of freedom. For one degree of freedom, the distribution has especially heavy tails—the variance is infinite—but as df grows, the distribution approaches a $\mathcal{N}(0,1)$.

by

$$\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \sim \sqrt{\frac{\chi^2_{n-1}}{n-1}}.$$

Then

$$\left(\frac{\overline{\mathbf{x}}-\mu}{\frac{\sigma}{\sqrt{n}}}\bigg/\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}\right) = \frac{\overline{\mathbf{x}}-\mu}{\hat{\sigma}/\sqrt{n}}.$$

The key stumbling block, the unknown value of σ, cancels out from the numerator and denominator. This is a work of genius by Mr. Student, because he could calculate the exact shape of the distribution through straightforward manipulation of the Normal and χ^2 tables.[3] Some t distributions for various degrees of freedom are pictured in Figure 9.4.

- The t_1 distribution (i.e., $n = 2$) is called a Cauchy distribution.
- As $n \to \infty$, the t_n distribution approaches the $\mathcal{N}(0,1)$ distribution.

[3] *Student* is actually Mr. William Sealy Gosset, who published the t distribution in 1908 based on his work as an employee of the Guinness Brewery.

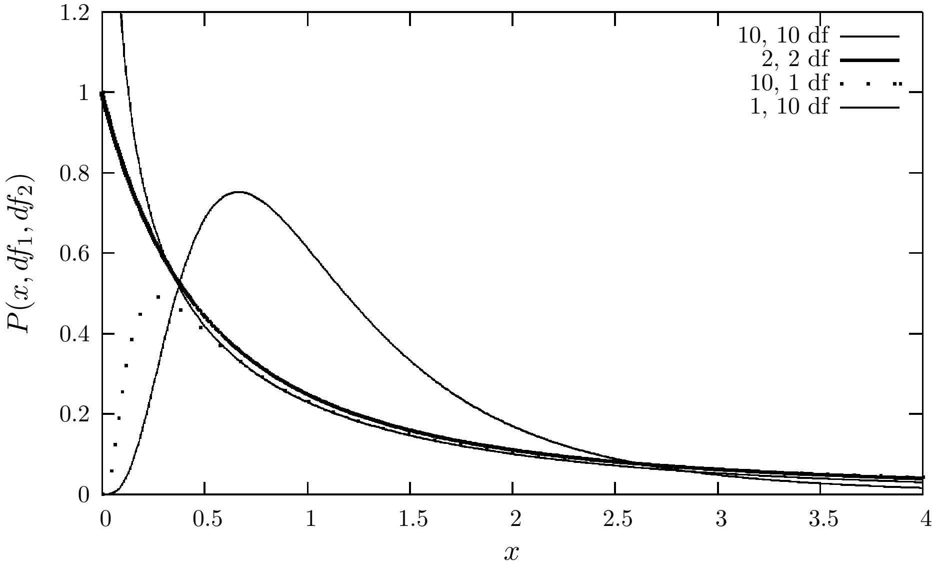


Figure 9.5 The F distribution for various pairs of numerator/denominator degrees of freedom.

F DISTRIBUTION Instead of a ratio of an $\mathcal{N}$ and a $\sqrt{\chi^2}$, you could also take the ratio of two χ^2-distributed variables. The σ's in both denominators would again cancel out, leaving a distribution that could be calculated from the χ^2 tables. This is the derivation and definition of the F distribution:

$$[\chi^2_m/m]/[\chi^2_n/n] \sim F(m, n).$$

See Figure 9.5 for some sample F distributions.

Also, consider the square of a t distribution. The numerator of a t_n distribution is a Normal distribution, so its square is a χ^2_1; the denominator is the square root of a χ^2_n distributed variable, so its square is a χ^2_n. Thus, the square of a t_n-distributed variable has an $F_{1,n}$ distribution as well.

The F distribution allows us to construct tests comparing one χ^2-distributed variable in the numerator to another χ^2-distributed variable in the denominator, and either of these χ^2 variables could be the sum of an arbitrary number of elements. We can thus use the F distribution to construct comparisons among relatively complex statistics.

LOOKUP TABLES There are three things that cover most of what you will be doing with a distribution: finding values of its PDF, values of its CDF, and occasionally values of its inverse CDF. For a single quick number, see the command-line program `apop_lookup`. For getting the the value of the PDFs at a given point from inside a program, here are the headers for the GSL's PDF lookup functions:

```
double gsl_ran_gaussian_pdf (double x, double sigma);
double gsl_ran_tdist_pdf (double x, double df);
double gsl_ran_chisq_pdf (double x, double df);
double gsl_ran_fdist_pdf (double x, double df1, double df2);
```

- The prefix `gsl_ran` indicates that these functions are from the random number generation module (`#include <gsl/gsl_randist.h>`). Random number generation itself will be delayed to Chapter 11.
- The mean for the Normal function is fixed at zero, so modify X accordingly, e.g., if X is drawn from a $\mathcal{N}(7, 1)$, then ask the GSL for `gsl_ran_gaussian_pdf(X-7, 1)`.

The next most common distribution calculation found in tables in the back of statistics texts is calculating the CDF above or below a point. The `P`-functions below calculate the CDF below a point, i.e. $\int_{-\infty}^{X} f(y)dy$, while the `Q`-functions calculate the CDF above a point, i.e. $\int_{X}^{\infty} f(y)dy$. These sum to one, so you can express any area in terms of whichever function is clearest.

Here is the list of functions:

```
double gsl_cdf_gaussian_P (double x, double sigma);
double gsl_cdf_tdist_P (double x, double df);
double gsl_cdf_chisq_P (double x, double df);
double gsl_cdf_fdist_P (double x, double df1, double df2);
```

. . . plus all of these with the `P` replaced by a `Q`.

These will be used to test *hypotheses*, which in this context are claims like $\mu > 0$. If you are shaky with hypothesis tests, see the next section. But if you are well-versed with the traditional notation for hypothesis tests, notice that the overuse of the letter P can easily lead to confusion. The `gsl_cdf_gaussian_P` function gives what is known as the p-value for the one-tailed test that the mean is less than zero, and the `gsl_cdf_gaussian_Q` function gives the infrequently-used q-value for the same hypothesis. Put another way, if we find that the mean of our Normally distributed data is 2.2 standard deviations below zero, then we reject the one-tailed hypothesis that the mean is less than or equal to zero with confidence `1 - gsl_cdf_gaussian_P(2.2, 1) == gsl_cdf_gaussian_Q(2.2, 1)`.

For a hypothesis that $\mu > 0$, everything is reversed. Here is a table that strives to clarify which function goes with the confidence with which the null is rejected and which goes with the p-value, and when:

	$H_0 : \mu > 0$	$H_0 : \mu < 0$
confidence	gsl_..._P	gsl_..._Q
p-value	gsl_..._Q	gsl_..._P

For a centered two-tailed test, the p-value takes the form

```
2 * GSL_MIN(gsl_ran_gaussian_P(mu, sigma), gsl_ran_gaussian_Q(mu, sigma))
// or equivalently,
2 * gsl_ran_gaussian_Q(fabs(mu), sigma)
```

The confidence with which we fail to reject the two-tailed null is one minus this.

In the other direction, we may want to know where we will need to be to reject a hypothesis with 95% certainty. For example, a value-at-risk oriented regulator will want to know the worst one-day loss a bank can expect over a month. To formalize the question, what is the value of the 1-in-20, or 5%, point on the CDF? Assuming a Normal(μ, σ) distribution of profit and loss,[4] the bank will report a value at risk of gsl_cdf_gaussian_Pinv (0.05, σ) + μ. Here are the requisite function declarations:

```
double gsl_cdf_gaussian_Pinv (double p, double sigma);
double gsl_cdf_tdist_Pinv (double p, double df);
double gsl_cdf_chisq_Pinv (double p, double df);
double gsl_cdf_fdist_Pinv (double p, double df1, double df2);
```

. . . plus all of these with the Pinvs replaced by Qinvs.

$\mathbb{Q}_{9.2}$ The *power* of a test is the likelihood of successfully rejecting the null hypothesis if there really is an effect in the data and the null should be rejected (see page 335 for more). When designing an experiment, you will need to estimate the power of a given design so you can decide whether to gather ten samples or a million.
I expect the mean of my Normally-distributed data to be 0.5, and $\hat{\sigma}$ to be 1.1. Given these assumptions, what must n be to reject the null hypothesis $\mu = 0$ with 99.9% confidence? What if the Normal approximation assumption is deemed inapplicable, so the data is taken to be t-distributed?

[4] This assumption is false. Securities typically have leptokurtic (fat-tailed) returns; see page 230.

$\sum$

- The square of a Normal distribution is a χ^2 distribution.
- Both of these distributions rely on an unknown variance σ^2. We can guess at σ^2, but then our confidence intervals are mere approximations as well.
- The ratio of a Normal over the square root of a transformed χ^2 distribution is a t distribution. By taking the ratio of the form of the two distributions, the unknown σ's cancel out, so a valid confidence region can be constructed from a finite amount of data.
- The ratio of two χ^2 distributions is an F distribution. Again, the unknown σ's cancel out.

9.3 TESTING A HYPOTHESIS

The chapter to this point has discussed how certain manners of gathering data and aggregating it into a statistic, such as taking its mean or taking the sum of squares, lead to certain known distributions. Thus, if we have a statistic produced in such a manner, we can evaluate the confidence with which a claim about that statistic is true. For example, the mean of a data set can be transformed to something having a t distribution (assuming the CLT holds). Similarly for the difference between the means for two data sets, so a precise probability can be placed on claims about the difference in two means.

CLAIMING A FIXED MEAN

This test is sometimes called a z-test, but see the footnote below. The claim is that the mean of a column of data is equal to μ_H. The procedure to the test:

- Find the mean of the data $\hat{\mu}$.
- Given the variance of the data set $\hat{\sigma}_d^2$, estimate the standard deviation of the mean via $\hat{\sigma}_m = \hat{\sigma}_d/\sqrt{n}$.
- For a one-tailed test, find the percentage of the t_{n-1} distribution that is over $|\mu_H - \hat{\mu}|/(\hat{\sigma}/\sqrt{n})$, i.e., `gsl_cdf_tdist_Q(fabs(`$\mu_H - \hat{\mu}$`)/`$\hat{\sigma}_m$`, n-1)`. Report this as the p-value.
- For a two-tailed test, report twice the calculated number as the p-value.

Q9.3

Can we reject the claim H_0: *The typical Puerto Rican county has over a 50% poverty rate*? Use the county-level info from the `poverty_pct_all` column from the `income` table of `data-census.db` as your data set. Given that the US Census Bureau defines poverty by the cost of living in the main part of the United States, how would you interpret your result?

Consider the parameters $\hat{\boldsymbol{\beta}}$ from the OLS regression, which you will recall takes the form $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. This is a more complex expression than a simple mean, but each element of the $\hat{\boldsymbol{\beta}}$ vector, $\hat{\beta}_0$, $\hat{\beta}_1$, …, can still be shown to have a simple t distribution. Thus, the standard test that a regression parameter β_i is significantly different from zero is a variant of the test here regarding a t-distributed scalar. The full details will be given in the section on regression tests below.

Comparing the mean of two sets

Among the most common and simplest questions is: are two sets of observations from the same process? Chapter 3 (page 109) already showed how to do a t test to test this claim.[5] You are encouraged to reread that section with an eye toward the test procedure.

Reporting confidence

There is some tradition of reporting only whether the p value of a test is greater than or less than some artificial threshold, such as $p > 0.05$ or $p < 0.05$. But Gigerenzer (2004) cites Fisher (1956) as stating:

> …no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.

Based on this observation, it would be better form to list the actual confidence calculated, and allow the reader to decide for herself whether the given value provides small or great confidence in the results. The error bars from Chapter 5 provide a convenient way to do this.

The paired t test is a common variant to the standard t test. Say that the data are paired in the sense that for each element in the first set, there is an element in the second set that is related; put another way, this means that for each element a_i, there is a corresponding element b_i such that the difference $a_i - b_i$ makes real-world sense. For example, we could look at student scores on a test before a set of lessons and scores by the same students after the lessons. Then, rather than looking at the t distribution for the *before* data and comparing it to the t distribution for

[5] There is no standardized naming scheme for tests. A test basically consists of three components:
1. the context,
2. the statistic to be calculated, and
3. the distribution that the statistic is compared to.
Thus, there are tests with names like the paired data test (using component 1), sum of squares test (component 2), or F test (component 3).

There is no correct way to name a procedure, but you are encouraged to avoid approach #3 where possible. First, there are really only about four distributions (Normal, χ^2, t, F) that are used in most real-world applications, which means that approach #3 gives us only four names for myriad tests. Two people could both be talking about *running a chi-squared test* and find that they are talking about entirely different contexts and statistics.

There is an odd anomaly regarding naming customs for the Normal distribution: rather than calling the statistic to be compared to the Normal a *normal statistic* or *Gaussian statistic*, it is typically called a z statistic. There is a z distribution, but it has nothing to do with the z test: it is one half the log of an F distribution, and is no longer commonly used because the F is slightly more convenient.

Finally, which distribution to use depends on the context and data: if a statistic has a t distribution for small n, then it approaches Normal as $n \to \infty$, so we could easily find ourselves in a situation where we are looking up the statistic for what is called a t test on the Normal distribution tables, or looking up a z statistic in the t distribution tables.

the *after* data, we could look at the vector of differences $a_i - b_i$ and find the confidence with which zero falls in the appropriate t distribution. This is generally a more powerful test, meaning that we are more likely to reject the null hypothesis of no difference between the two vectors, and therefore the paired t test is generally preferable over the unpaired version (when it is applicable). Apophenia provides the `apop_paired_t_test` function to run this test where appropriate.

χ^2-BASED TESTS One quick application of the χ^2 distribution is for testing whether a variance takes on a certain value. We posit that the denominator of Equation 9.2.1 is a fixed number, and then check the χ^2 tables for the given degrees of freedom. This is a relatively rare use of the distribution.

A more common use is to take advantage of the summation property to combine individual statistics into more elaborate tests. Any time we have statistics of the form (observed $-$ expected)^2/expected, where (observed $-$ expected) should be Normally distributed, we can use Pearson's Taylor series approximation to piece together the statistics to form a χ^2 test. There are examples of this form in the section on ANOVA and goodness-of-fit testing below.

F-BASED TESTS Because of all the squaring that goes into a χ^2 distributed statistic, x and $-x$ are indistinguishable, and so it becomes difficult to test one-tailed claims of the form $a > b$. We could use the t test for a one-tailed claim about a single variable, or an F test for a combination of multiple variables.

Let H_0 be the claim that $\mathbf{Q}'\boldsymbol{\beta} = \mathbf{c}$. This is a surprisingly versatile hypothesis. For example, say that $\boldsymbol{\beta}$ is a vector with three elements, $\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}$, $\mathbf{Q} = \begin{bmatrix}1\\0\\0\end{bmatrix}$, and $\mathbf{c} = [7]$. Then H_0 is $\beta_1 = 7$. Or, $\mathbf{Q} = \begin{bmatrix}1\\-1\\0\end{bmatrix}$ and $\mathbf{c} = \begin{bmatrix}0\end{bmatrix}$ gives $H_0 : \beta_1 = \beta_2$. Or, say we want to test $H_0 : \beta_2 = 2\beta_3$. Then let $\mathbf{Q} = \begin{bmatrix}0\\1\\-2\end{bmatrix}$ and $\mathbf{c} = 0$. In ANOVA terminology, a hypothesis about a linear combination of coefficients is known as a *contrast*.

To test all three hypotheses at once, simply stack them, one hypothesis to a row:

$$\mathbf{Q}' = \begin{bmatrix}1 & 0 & 0\\ 1 & -1 & 0\\ 0 & 1 & -2\end{bmatrix} \qquad \mathbf{c} = \begin{bmatrix}7\\0\\0\end{bmatrix}. \tag{9.3.1}$$

Any linear (or affine) hypothesis having to do with a parameter β can be fit into this form.

Define q to be the number of constraints (rows of $\mathbf{Q}'$), n the sample size, and k the number of parameters to be estimated ($\boldsymbol{\beta}$). As before, let $\underline{\mathbf{X}}$ represent $\mathbf{X}$ normalized so that each column but the first has mean zero, and the first column is the constant vector $\mathbf{1}$. Now, if H_0 is true and $\boldsymbol{\beta}$ was estimated using OLS, then $\mathbf{Q}'\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{c}, \sigma^2\mathbf{Q}'(\underline{\mathbf{X}}'\underline{\mathbf{X}})^{-1}\mathbf{Q})$,[6] and we can construct a χ^2-distributed linear combination of the square of q standard Normals via

$$\frac{(\mathbf{Q}'\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{Q}'(\underline{\mathbf{X}}'\underline{\mathbf{X}})^{-1}\mathbf{Q}]^{-1}(\mathbf{Q}'\hat{\boldsymbol{\beta}} - \mathbf{c})}{\sigma^2} \sim \chi^2_q. \tag{9.3.2}$$

Alternatively, say that we are testing the value of the variance of the regression error, and $\boldsymbol{\epsilon} \sim \mathcal{N}$; then

$$\frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{\sigma^2} \sim \chi^2_{n-k}. \tag{9.3.3}$$

As above, we can divide scaled versions of Equation 9.3.2 by Equation 9.3.3 to give us a statistic with an F distribution and no unknown σ^2 element:

$$\frac{n-k}{q}\frac{(\mathbf{Q}'\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{Q}'(\underline{\mathbf{X}}'\underline{\mathbf{X}})^{-1}\mathbf{Q}]^{-1}(\mathbf{Q}'\hat{\boldsymbol{\beta}} - \mathbf{c})}{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}} \sim F_{q,n-k}. \tag{9.3.4}$$

If you have read this far, you know how to code all of the operations in Equation 9.3.4. But fortunately, Apophenia includes a function that will calculate it for you. To do this, you will need to feed the function an estimate of $\boldsymbol{\beta}$ and an `apop_data` set indicating the set of contrasts you wish to test, whose vector element is $\mathbf{c}$ and matrix element is $\mathbf{Q}'$. As in Equation 9.3.1, each *row* of the input matrix represents a hypothesis, so to test all three equality constraints at once, you could use a vector/matrix pair like this:

```
double line = {
    7, 1, 0, 0
    0, 1, -1, 0
    0, 0, 1, -2};
apop_data *constr = apop_line_to_data(line, 3, 3, 3);
```

[6] For any other method, the form of the variance is $\mathbf{Q}'$(the variance from Section 8.2)$\mathbf{Q}$. See, e.g., Amemiya (1994).

The final [vector|matrix] form for the constraint matches the form used for constraints on pp 152–153, but in this case the constraints are all equalities.

Listing 9.6 presents a full example.

- It runs a regression on the males-per-female data from page 267, so link this code together with that code.
- The constraint is only one condition: that $\beta_3 = 0$.
- The `apop_F_test` function takes in a set of regression results, because the F test as commonly used is so closely married to OLS-type regressions.

```
#include "eigenbox.h"

int main(){
    double line[] = {0, 0, 0, 1};
    apop_data *constr = apop_line_to_data(line, 1, 1, 3);
    apop_data *d = query_data();
    apop_model *est = apop_estimate(d, apop_ols);
    apop_model_show(est);
    apop_data_show(apop_f_test(est, constr));
}
```

Listing 9.6 Run an F-test on an already-run regression. Online source: `ftest.c`.

Here is a useful simplification. Let R^2 be the coefficient of determination (defined further below), n be the number of data points, k be the number of parameters (including the parameter for the constant term), and ϕ be the F-statistic based on $\mathbf{Q} = \mathbf{I}$ and $\mathbf{c} = \mathbf{0}$. Then it can be shown that

$$\frac{(n-k)R^2}{k(1-R^2)} = \phi. \tag{9.3.5}$$

$\mathbb{Q}_{9.4}$

Verify the identity of Equation 9.3.5 using Equation 9.3.4 and these definitions (from page 228):

$$\begin{aligned}
\mathbf{y}_{\text{est}} &\equiv \mathbf{X}\boldsymbol{\beta} \text{ (the estimated value of } \mathbf{y}\text{)},\\
SSR &\equiv \sum (\mathbf{y}_{\text{est}} - \overline{\mathbf{y}})^2,\\
SSE &\equiv \boldsymbol{\epsilon}'\boldsymbol{\epsilon}, \text{ and}\\
R^2 &\equiv \frac{SSR}{SSE}.
\end{aligned}$$

Statistical custom is based on the availability of computational shortcuts, so the F statistic of Equation 9.3.5 often appears in the default output of many regression packages.[7] It is up to you to decide whether this particular test statistic is relevant for the situation you are dealing with, but because it is a custom to report it, Apophenia facilitates this hypothesis test by assuming it as the default when you send in `NULL` variables, as in `apop_F_test(estimate, NULL)`.

$\mathbb{Q}_{9.5}$ Verify the identity of Equation 9.3.5 by running a linear regression on the data set you produced for the exercise on page 278, then passing the `apop_model` thus produced to `apop_F_test` to find the F statistic and Apohenia's R^2-finding function to find the SSE and SSR.

$\sum$

- ➤ The simplest hypothesis test regarding the parameters of a model is the t test. It claims that the mean of a data set is different from a given value of μ. A special case is the claim that the mean of two data sets differ.
- ➤ The χ^2 test allows the comparison of linear combinations of allegedly Normal parameters. But since everything is squared to get the χ^2 parameter, it can not test asymmetric one-tailed hypotheses.
- ➤ The F test provides full generality, and can test both one-tailed and two-tailed hypotheses, and claims that several contrasts are simultaneously true.

9.4 ANOVA

ANOVA is a contraction for *analysis of variance*, and is a catch-all term for a variety of methods that aim to decompose a data set's variance into subcategories. Given a few variables and a few groups, is there more variation between groups or within groups? Can the variation in a dependent variable be attributed primarily to some independent variables more than others?

The descriptive portion of ANOVA techniques was covered back on pages 224–227. This section covers the hypothesis testing part of ANOVA.

You may want to re-run `metroanova.c`, which first appeared as Listing 7.2 on page 226. It produces an ANOVA table that includes several sums of squared errors, and the ratio between them. At this point, you will recognize a sum of squared errors as having a χ^2 distribution (assuming the errors are Normally distributed), and the df-weighted ratio of two sums of squared errors as being F-distributed.

[7] Scheffé (1959) parodies the singular focus on this form by calling it *"the" F test* throughout the book.

Thus, the traditional ANOVA table includes an F test testing the claim that the among-group variation is larger than the within-group variation, meaning that the grouping explains a more-than-random amount of variation in the data.

Independence

The crosstab represents another form of grouping, where the rows divide observations into the categories of one grouping, and the columns divide the observations into categories of another. Are the two groupings independent?

To give a concrete example, say that we have a two-by-two array of events, wherein 178 people chose between up/down and left/right:

	Left	Right	Σ
Up	30	86	116
Down	24	38	62
Σ	54	124	178

Is the incidence of Up/Down correlated to the incidence of Left/Right, or are the two independent? Draws from the four boxes should follow a Multinomial distribution: if Up/Down were a Bernoulli draw with probabilities p_U and p_D, and Left/Right were a separate, independent Bernoulli draw with probabilities p_L and p_R, then the expected value of Up/Left would be $E_{UL} = np_U p_L$, and similarly for E_{DL}, E_{UR}, and E_{DR}. Notating the actual incidence of Up/Left as $O_{UL} = 30$, we can use the fact (from page 301) that the χ^2 is a reasonable approximation of errors from a Multinomial distribution to say that the observed variance over the expected value $(O_{UL} - E_{UL})^2/E_{UL} \sim \chi^2$. Similarly for the other three cells, so the sum

$$\frac{(O_{UL} - E_{UL})^2}{E_{UL}} + \frac{(O_{UR} - E_{UR})^2}{E_{UR}} + \frac{(O_{DL} - E_{DL})^2}{E_{DL}} + \frac{(O_{DR} - E_{DR})^2}{E_{DR}} \sim \chi^2_1. \tag{9.4.1}$$

This expression has one degree of freedom because the horizontal set has two elements and one mean ⇒ one df; similarly for the vertical set; and 1 df × 1 df = 1 df. If there were three rows and six columns, there would be $2 \times 5 = 10$ df.

Listing 9.7 calculates this, once the long way and twice the short way. The `calc_chi_squared` function calculates Equation 9.4.1, using the `one_chi_sq` function to calculate each individual term. Finally, `main` gathers the data and calls the above functions. After all that, it also calls `apop_test_anova_independence`, which does all this work for you on one line.

The distribution of means of a series of Binomial draws will approach a Normal as $n \to \infty$, but for many situations, n is closer to around ten. For such a case, Fisher

```
#include <apop.h>

double one_chi_sq(apop_data *d, int row, int col, int n){
    APOP_ROW(d, row, vr);
    APOP_COL(d, col, vc);
    double rowexp = apop_vector_sum(vr)/n;
    double colexp = apop_vector_sum(vc)/n;
    double observed = apop_data_get(d, row, col);
    double expected = n * rowexp * colexp;
    return gsl_pow_2(observed - expected)/expected;
}

double calc_chi_squared(apop_data *d){
  double total = 0;
  int n = apop_matrix_sum(d->matrix);
    for (int row=0; row <d->matrix->size1; row++)
        for (int col=0; col <d->matrix->size2; col++)
            total += one_chi_sq(d, row, col, n);
    return total;
}

int main(){
  double dataline[] = { 30,86,
                        24,38 };
  apop_data *data = apop_line_to_data(dataline, 0, 2, 2);
  double stat, chisq;
    stat = calc_chi_squared(data);
    chisq = gsl_cdf_chisq_Q(stat, (data->matrix->size1 - 1)* (data->matrix->size2 - 1));
    printf("chi squared statistic: %g; p, Chi-squared: %g\n", stat,chisq);
    apop_data_show(apop_test_anova_independence(data));
    apop_data_show(apop_test_fisher_exact(data));
}
```

Listing 9.7 Pearson's χ^2 test and Fisher's Exact test. Online source: `fisher.c`.

(1922) calculated the probability of a given table using direct combinatorial computation. The equations for the Fisher exact test are a mess, but the story is the same as above—its null hypothesis is that Up/Down and Left/Right are independent—and its calculation is trivial: just call `apop_test_fisher_exact`, as in the last line of Listing 9.7.

※ *Scaling* How would the calculation be affected if we replicated every count in the data set into k counts, so $O'_{UL} = kO_{UL}$ and $E'_{UL} = kE_{UL}$? Then $(O'_{UL} - E'_{UL})^2/E'_{UL} = k(O_{UL} - E_{UL})^2/E_{UL}$. That is, scaling the data set by k scales the test statistic by k as well. For almost any data set, there exists a k for which the null hypothesis will be rejected.

Across data sets, the scale can easily be different, and statistical significance will be easier to achieve in the set with the larger scale. Generally, it is tenuous to assert that a data set with a test statistic in the 96th percentile of the χ^2 distribution diverges from independence less than a data set whose test statistic is in the 99.9th percentile. Use the test to establish whether the data rejects the null hypothesis, then use other methods (a simple covariance will often do) to establish the magnitude of the difference.

For comparison, prior tests involving the mean are not as sensitive to scale. Notably, consider the ratio upon which the Central Limit Theorem is based, after every element of the data vector $\mathbf{x}$ is scaled by k:

$$\begin{aligned}\frac{\text{mean}}{\sqrt{\text{var}/n}} &= \frac{\sum(kx - k\overline{\mathbf{x}})/n}{\sqrt{\sum(kx - k\overline{\mathbf{x}})^2/n^2}}\\ &= \frac{\sum(x - \overline{\mathbf{x}})}{\sqrt{\sum(x - \overline{\mathbf{x}})^2}}.\end{aligned}$$

All else equal, the ratio of the mean to $\sqrt{\hat{\sigma}^2/n}$ (often written $\hat{\sigma}/\sqrt{n}$) is not affected by the scale of $\mathbf{x}$, or even the number of elements in the data set, the way the χ^2 statistic above was affected by rescaling.

9.5 REGRESSION

In the prior chapter, we used the linear regression model for purely descriptive purposes, to find the best way to project $\mathbf{y}$ onto $\mathbf{X}$. If we add the assumption that $\boldsymbol{\epsilon}$ is Normally distributed, then we can also test hypotheses regarding the parameter estimates. Given this assumption, it can be shown that the coefficients on the independent variables (the $\boldsymbol{\beta}$ vector) have a t distribution, and therefore the confidence with which an element of $\boldsymbol{\beta}$ differs from any constant can be calculated (see, e.g., Greene (1990, p 158)).

The covariance matrix of $\boldsymbol{\beta}_{\text{OLS}}$ is $\Sigma = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, where σ^2 is the variance of the regression's error term: if $\boldsymbol{\epsilon}$ is the vector of errors, and there are n data points and k regressors (including any constant column $\mathbf{1}$), then $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}/(n-k)$ provides an unbiased estimate of σ^2.[8] The estimated variance of β_1 is the first diagonal element, Σ_{11}; the estimated variance of β_2 is the second diagonal element, Σ_{22}; and so on for all β_i.

As is typical for a test of a statistic of the data, the count of degrees of freedom is data points minus constraints; specifically, for n data points and k regression parameters (including the one attached to the constant column), $df = n - k$.

[8] The form of the variance of the error term analogizes directly with the basic one-variable unbiased estimate of variance, $\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2/(n-1)$. First, the setup of OLS guarantees that $\overline{\boldsymbol{\epsilon}} = 0$, so $\epsilon_i - \overline{\boldsymbol{\epsilon}} = \epsilon_i$, and thus $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ matches the numerator in the basic formula. The denominator in all cases is the degrees of freedom; for example, with k regressors and n data points, there are $n-k$ degrees of freedom.

Given the estimated variance $\hat{\sigma}^2$ for β_i and any constant c, we could write down a test statistic $|\beta_i - c|/\hat{\sigma}$, and then check that statistic against the t_{n-k} distribution to test the claim that $\beta_i = c$. This test bears a close resemblance to the test for the mean of a data set (also a t-distributed scalar statistic) presented on page 307. If you have a joint hypotheses about contrasts among the elements of $\boldsymbol{\beta}$, you can directly apply the above discussion of F tests: just use the estimated mean of $\boldsymbol{\beta}$, its covariance matrix Σ, and $n - k$ degrees of freedom.

Comparison with ANOVA If a regression consists of nothing but dummy variables, then it can be shown that the process is equivalent to the ANOVA-style categorization methods above.

$\mathbb{Q}_{9.6}$ Alaska is famously low on females, due to its many features that distinguish it from the lower 48 states. Create a dummy variable where 1=Alaska, 0=other state, and regress males per female against both the Alaska dummy and population density (and a constant $\mathbf{1}$, of course). Are one or both of the independent variables significant?

$\mathbb{Q}_{9.7}$ Run an independence test on the two-by-two table whose row categories are Alaska and not-Alaska, and whose column categories are males and females. (*Hint*: you will need to combine the population and males per females columns to produce a count for each region, then sum over all regions.)
How does the test differ if you compare the percent male/female or the total count of males and females in each region? What changes in the story underlying the test, and which version better represents the hypothesis?

OLS (along with its friends) has two real advantages over testing via crosstab approaches like ANOVA. First, it readily handles continuous variables, which ANOVA can handle only via approximation by rough categories.

Second, it allows the comparison of a vast number of variables. ANOVAs typically top out at comparing two independent variables against one dependent, but an OLS regression could project the dependent variable into a space of literally hundreds of independent variables. In fact, if you run such a regression, you are basically guaranteed that some number of those variables will be significant.

The multiple testing problem Freedman (1983) showed the dangers of *data snooping* by randomly generating a set of 51 columns of 100 random numbers each.[9] He set one column to be the dependent variable to be ex-

[9] Data snooping used to also be called *data mining*, but that term has lost this meaning, and is now used to refer to categorization techniques such as classification trees.

plained, and the other fifty to be potential regressors. Using a simple exploratory technique, he culled the fifty potential explanatory variables down to 15 variables. He then ran a 15-variable regression, and found that 14 variables were significant with a p-value better than 25%, and six were significant with p better than 5%. Other tests of the regression also indicated a good fit. But the data was pure noise by construction.

Recall from the first paragraph of this book that there are two goals of statistical analysis, and they directly conflict. If a researcher spends too much time looking for descriptive statistics about the data, then he is committing informal data snooping, akin to Freedman's initial exploratory regression, and thus biases the chances of rejecting a null in her favor. But it would be folly for the researcher to not check the data for outliers or other quirks before running the regression, or to embark upon producing an entirely new data set for every regression.

What is the correct balance? Statistics has no answer to this, though most statisticians do. Those in the descriptive-oriented camp are very serious about the importance of good graphical displays and viewing the data every way possible, while those in the testing-oriented camp believe that so much pre-test searching is simply asking for apophenia.

Here is another spin on the issue: people who are testing exactly one hypothesis tend to develop an affection for it, and become reluctant to reject their pet hypothesis. Thus, research as conducted by humans may improve if there are multiple hypotheses simultaneously competing. Chamberlin (1890, p 93) explains:

> Love was long since represented as blind, and what is true in the personal realm is measurably true in the intellectual realm. . . . The moment one has offered an original explanation for a phenomenon which seems satisfactory, that moment affection for his intellectual child springs into existence; and as the explanation grows into a definite theory, his parental affections cluster about his intellectual offspring, and it grows more and more dear to him, so that, while he holds it seemingly tentative, it is still lovingly tentative. . . . The mind lingers with pleasure upon the facts that fall happily into the embrace of the theory, and feels a natural coldness toward those that seem refractory. . . . There springs up, also, an unconscious . . . pressing of the facts to make them fit the theory. . . . The search for facts, the observation of phenomena and their interpretation, are all dominated by affection for the favored theory until it appears to its author or its advocate to have been overwhelmingly established. The theory then rapidly rises to the ruling position, and investigation, observation, and interpretation are controlled and directed by it. From an unduly favored child, it readily becomes master, and leads its author whithersoever it will.

His solution, as above, is to test multiple hypotheses simultaneously. "The investigator thus becomes the parent of a family of hypotheses; and, by his parental relation to all, he is forbidden to fasten his affections unduly upon any one." He also points out that maintaining multiple hypotheses allows for complex explanations about how an outcome was partly caused by one factor, partly by another, partly by another. After all, Nature is not compelled to conform to exactly one hypothesis.

Apophenia's model-as-object makes it very easy to test or mix diverse hypotheses, as per Chamberlin's suggestion, and you will see more methods of comparing models in later chapters. But as the number of models grows, the odds of failing to reject at least one model purely by chance grows as well. There is no hard-and fast rule for determining the "correct" number of models to test; just bear in mind that there is a tension among multiple goals and a balance to be struck between them.

Correcting for multiple testing Moving on from informally poking at the data, consider the case when the experiment's basic design involves a systematic, fixed series of tests, like running a separate test for every genetic marker among a list of a million. This is known as the *multiple testing problem*, and there is a simple means of correcting for it.

Say that a number is drawn from $[0, 1]$, and the draw is less than p with probability p. Then the probability that a draw is greater than p is $1 - p$, and the probability that n independent draws are all greater than p is $(1 - p)^n$, which can be reversed to say that the probability that at least one of n independent draws is less than p is $1 - (1 - p)^n$.

Thus, the probability that, out of n tests with a fixed p-value, at least one will indicate a positive result is $\rho = 1 - (1 - p)^n$. For example, with $p = 0.05$ and $n = 100$, the likelihood of rejecting the null at least once is $1 - (1 - 0.05)^{100} \approx 99.4\%$.

We can instead fix ρ at a value like 0.05 or 0.01, and reverse the above equation to find the p-value for the individual tests that would lead to rejection of all nulls with 5% or 1% likelihood. A line or two of algebra will show that $p = 1 - (1 - \rho)^{1/n}$. For $n = 100$ and $\rho = 0.05$, you would need to set the p-value for the individual tests to 0.0005128. In the example of testing $n = 1,000,000$ genetic markers, if the desired overall $\rho = 0.05$, then the p-value for the individual tests would be $5.129\text{e}{-8}$.

There is a wonderfully simple approximation for the above expression: just let $p = \rho/n$. For the first example above, this approximation is $0.05/100 = 0.0005$; for the second it is $0.05/1,000,000 = 5\text{e}{-8}$. Both of these approximations are within about $\pm 2.5\%$ of the true value.

Thus, we have a simple rule, known as the *Bonferroni correction*, for multiple tests: just divide the desired overall p-value by the number of tests to get the appropriate individual p-values. The correction is standard in biometrics but virtually unknown in the social sciences. When reading papers with pages of regressions and no corrections for multiple testing, you can easily apply this equation in your head, by multiplying the reported individual p-value by the number of tests and comparing that larger figure to the usual significance levels of 0.05 or 0.01.

- Because we know their expected mean and covariances, the regression parameters for OLS, IV, WLS, and other such models can be tested individually using the standard t test, or tested as a group via an F test.
- When running a battery of several tests (based on a regression or otherwise), use the Bonferroni correction to create a more stringent significance level. The common form of calculating the more stringent p-value is to simply divide the one-test p-value by the number of tests.

9.6 Goodness of fit

This section will present various ways to test claims of the form *the empirical distribution of the data matches a certain theoretical distribution*. For example, we often want to check that the errors from a regression are reasonably close to a Normal distribution.

The visually appealing way to compare two distributions is the Q–Q plot, which stands for quantile–quantile plot. The first (x, y) coordinate plotted is x_1 = the first percentile of your data and y_1 = the first percentile of the distribution you are testing, the second is x_2 = the second percentile of your data and y_2 = the second percentile of the ideal distribution, et cetera. To the extent that the data fits the ideal distribution, the points will draw out the $x = y$ line, while digressions from the line will stand out.

The first half of Listing 9.9 presents an example, displaying a plot to check whether precipitation is Normally distributed. It gathers the data in the usual `apop_query_to_data` manner, estimates the closest-fitting Normal distribution, and plots the percentiles of the data against the percentiles of the just-estimated distribution. As Figure 9.8 shows, the data set closely fits the Normal distribution (though the extremes of the bottom tail is a bit more elongated and the extreme of the top tail a bit less so).

Q$_{9.8}$ Modify Listing 9.9 to test whether temperature or log of temperature is Normally distributed. Would any other distribution fit better?

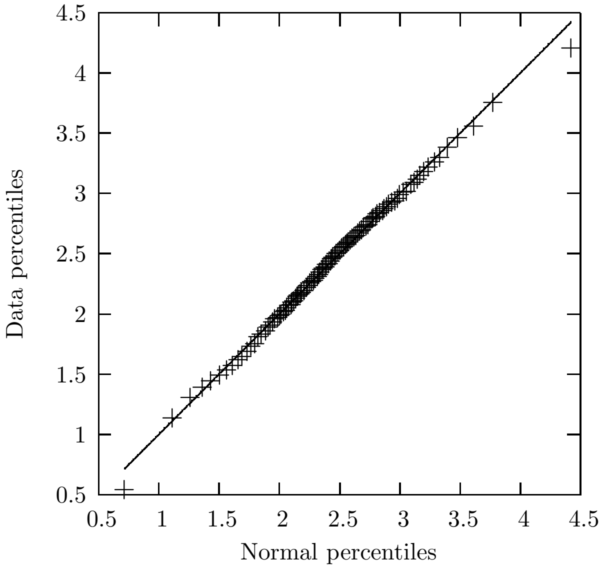


Figure 9.8 Percentiles of precipitation on the y axis plotted against percentiles of the Normal distribution along the x axis.

```
#include <apop.h>

int main(){
    apop_db_open("data-climate.db");
    apop_data *precip = apop_query_to_data("select PCP from precip");
    apop_model *est = apop_estimate(precip, apop_normal);
    Apop_col_t(precip, "PCP", v);
    apop_plot_qq(v, *est, "outfile.gnuplot");

    double var = apop_vector_var(v);
    double skew = apop_vector_skew(v)/pow(var, 3/2);
    double kurt = apop_vector_kurtosis(v)/gsl_pow_2(var) - 3;
    double statistic = v->size * (gsl_pow_2(skew)/6. + gsl_pow_2(kurt)/24.);
    printf("The skew is %g, the normalized kurosis is %g, "
            "and we reject the null that your data is Normal with %g confidence.\n",
        skew, kurt, gsl_cdf_chisq_P(statistic, 2));
}
```

Listing 9.9 Pull data; estimate the Normal that best fits the data; plot the data against the ideal distribution. The output is presented in Figure 9.8. Online source: `qqplot.c`.

Higher moments A slightly more rigorous alternative means of testing for Normality is to check the higher moments of the Normal distribution (Bowman & Shenton, 1975; Kmenta, 1986, pp 266–267). There is a more general chi-squared goodness-of-fit test for any distribution below.

A Normal distribution has only two parameters, the mean and the standard deviation, and everything else about the Normal distribution is defined therefrom. Notably, the third moment is zero, and the fourth moment is $3\sigma^4$.

We already have everything we need to calculate the distribution of these statistics. The skew and kurtosis are both the mean of an iid process (recall their definitions on page 230: a sum divided by n), so their square is $\sim \chi^2$. Let s be the third moment of the data divided by σ^3 and let κ be the fourth moment divided by σ^4. Then

$$L_s = n\left[\frac{s^2}{6}\right]$$

has a χ^2 distribution with one degree of freedom, as does

$$L_k = n\left[\frac{(\kappa-3)^2}{24}\right].$$

Some prefer to test both simultaneously using

$$L_{sk} = n\left[\frac{s^2}{6}+\frac{(\kappa-3)^2}{24}\right],$$

which has a χ^2 distribution with two degrees of freedom.

The second half of Listing 9.9 translates this formula into code. Given the Q–Q plot, it is no surprise that the test soundly fails to reject the null hypothesis that the data is Normally distributed.

Another alternative, keeping with the theme of this book, would be to bootstrap the variance of the kurtosis, which would let you find a confidence interval around $3\sigma^4$ and state with some percentage of certainty that the kurtosis is or is not where it should be; this suggestion is put into practice on page 365.

CHI-SQUARED GOODNESS-OF-FIT TEST Say that we have a histogram and a vector of points that we claim was drawn from that histogram. It would be nice to test the confidence with which our claim holds; this is a goodness-of-fit test.

Mathematically, it is simple. We have k bins, and two histograms: `h0` holds the histogram from which the draws were allegedly made, and `h1` holds the data.[10] Then

[10] Recall from page 314 that scaling matters for a χ^2 test: the histograms representing two PDFs will each sum to one (by definition), while a histogram representing the density of a population of size n will have bins summing to n (by definition). That means that the χ^2 statistics for a test of the PDFs and a test of the distribution of counts will be different, with the null more likely to be rejected for the distribution of counts. So when investigating a histogram, be careful that you are testing the right hypothesis; claims about the distribution of a population are typically best represented by a test of the PDFs ($\Sigma = 1$) rather than the counts ($\Sigma = n$).

$$\sum_{i=0}^{k} \frac{(\texttt{h0->\ bins[i]} - \texttt{h1->\ bins[i]})^2}{\texttt{h0->\ bins[i]}} \sim \chi^2_{k-1}. \tag{9.6.1}$$

You will recognize this form as matching the *(observed - expected)*^{2}*/expected* form from the ANOVA tests earlier in this chapter.

Q9.9

On page 173, you plotted the leading digit of an arbitrary data set, and saw that it sloped sharply down. Now use a chi-squared goodness of fit test to formally check that your data set fits Equation 5.4.1.

- Write a function to produce a vector of nine elements, with the count of elements in each slot equal to the number of data points with the given leading digit. Don't forget that vectors count from zero but you want the first slot to represent leading digit one, and to rescale your final vector so that it is a PMF (i.e., its elements sum to one).
- Equation 5.4.1 isn't quite a PMF: the sum of its values from one to nine isn't one. Thus, you will need to get the total mass, and rescale your calculations from Benford's equation accordingly.
- Once you have two nine-element vectors of equal mass, you can directly apply Expression 9.6.1 to find the χ^2 statistic and run the χ^2 test.

```
 1  #include <apop.h>
 2
 3  int main(){
 4      apop_db_open("data-climate.db");
 5      apop_data *precip = apop_query_to_data("select PCP from precip");
 6      apop_model *est = apop_estimate(precip, apop_normal);
 7      gsl_rng *r = apop_rng_alloc(0);
 8      apop_model *datahist = apop_estimate(precip, apop_histogram);
 9      apop_model *modelhist = apop_histogram_model_reset(datahist, est, 1e6, r);
10      apop_data_show(apop_histograms_test_goodness_of_fit(datahist, modelhist));
11  }
```

Listing 9.10 The same precipitation data, another test. Online source: `goodfit.c`.

Listing 9.10 tests whether the precipitation data is Normally distributed using the χ^2 goodness-of-fit test.

- Lines 1–6 are a repeat of the query and estimation from `qqplot.c` (page 320).

- Line eight turns the input data into a histogram. Notice that it uses the same apop_estimate form as other models, because a histogram is just another means of expressing a model.
- You can't do a goodness-of-fit test on just any two histograms: the bins have to match, in the sense that the range of each bin in the first histogram exactly matches the range in the corresponding bin of the second. The easiest way to ensure that two histograms match is to generate the second histogram using the first as a template, which is what apop_histogram_model_reset does. If we wanted to compare two vectors via this test, this line would use apop_histogram_vector_reset.
- The new histogram gets filled via random draws from the model, which means that we need to give apop_histogram_model_reset the number of draws to make (here, 1e6), and a gsl_rng to provide random numbers. The use of the gsl_rng is covered in full on page 357.
- By line ten, we have two histograms representing the data and the model, and they are in sync. Thus, it is a simple matter to send the histograms to apop_histograms_test_goodness_of_fit to calculate the statistic in Expression 9.6.1.

KOLMOGOROV'S METHOD Kolmogorov (1933) suggested considering the steps in a histogram to be a Poisson process, and developed a test based upon this parametrization [see also Conover (1980)]. Given two histograms produced using one of the above-mentioned methods, apop_test_kolmogorov finds the maximum difference in the CMFs and find the probability that such a CMF would occur if both histograms were from the same base data set. Because this test uses the ordering of the slices of the PMF, while the Chi-squared test does not, this test generally has higher power.

Kolmogorov's test serves as yet another test for Normality, because it can compare a data set's CDF to that of a Normal distribution.

$\mathbb{Q}_{9.10}$

Is GDP per capita log-normally distributed?

- Pull the log of GDP per capita data from the data-wb.db data set.
- Create a histogram (i.e., estimate an apop_histogram model).
- Fit a Normal distribution and use it to create a matching histogram using apop_histogram_model_reset.
- Send the two histograms to apop_test_kolmogorov.
- How do the test results compare with those produced by apop_histograms_test_goodness_of_fit?

$\mathbb{Q}_{9.11}$ How about precipitation? Figure 9.8 gave the strong suggestion that the data is Normally distributed; modify the code from the last example to formally test the hypothesis that the data set is drawn from a Normal distribution using the Kolmogorov–Smirnov method.

$\sum$

- The Q–Q (quantile-to-quantile) plot gives a quick visual impression of whether the distribution of the data matches that of a theoretical distribution.
- We can test the claim that a data set is Normally distributed using the fact that the skew and kurtosis of a Normal distribution are fixed (given the mean and variance).
- More generally, we can compare any two distributions by dividing them into bins and comparing the square of the deviation of one distribution from another via a χ^2 test.
- The Kolmogorov–Smirnov test offers another method for comparing two distributions, which typically has more power than the Chi-squared method.