This document is an excerpt from
Resampling Stats in MATLAB
Daniel T. Kaplan
Copyright (c) 1999 by Daniel T. Kaplan, All Rights Reserved
This document differs from the published book in pagination and in the omission (unintentional, but unavoidable for technical reasons) of figures and cross-references from the book. It is provided as a courtesy to those who wish to examine the book, but not intended as a replacement for the published book, which is available from
Resampling Stats, Inc.
`www.resample.com`
703-522-2713

Chapter 3: Testing Hypotheses with Data

3.1 Concepts of Hypothesis Testing

In the first 3 games of the 1988 NBA playoff series between Boston and Detroit, Larry Bird scored 20 of 57 shots. Since Bird had previously made 48% of his shots, he would have been expected to score 28 out of the 57 attempts. The Washington Post said

> "Larry Bird is so cold he couldn't throw a beachball in the ocean ... They fully expect Bird to come out of his horrendous shooting slump." (May 30, 1988, p.D4)

The use of the word "slump" suggests that Bird is playing much worse than usual. In the case of a basketball game, there is little to be done except perhaps pep talks or — unimaginably — replacing Bird. But problems such as this are common in many settings. For example, in industrial production the quality of the goods produced may fall below the standards for a time and it is important to know whether this is just a chance fluctuation or whether the production line ought to be stopped to allow readjustments, whether new suppliers of raw materials ought to be sought, etc. Such problems are sometimes called *quality control* problems, and are an example of a common situation in statistical inference.

The question that statistical inference addresses is this: "Do the data actually support the conclusion that something has gone wrong?" In the case of Larry Bird, we want to know if he is actually in a slump. The 57 attempts made by Bird in the playoffs are a sample of all of the possible attempts by Bird. We would not expect him to make exactly 48% of his attempts even if he were playing perfectly up to par. Instead, he might make more or fewer scores than the average of 28 out of 57. How likely is it that Bird would score only 20 or fewer out of 57 shots if he were playing normally?

Example 1: A Basketball Slump?

We can set up the computer to simulate a player who makes, on average, 48% of shots. We then have the computer make 57 attempts and count how many successes occurred. This is done with the following commands:

Give a 48 percent chance of scoring a shot (coded as 1).

```
≫   bird = [48 1; 52 0];
```

Simulate 57 shots.

```
≫   a = sample(57, bird);
```

Count the number of baskets in the 57 shorts.

```
≫   count(a==1)
```

Due to sampling variability, the result you get by repeating these commands is probably not exactly 28 (the expected number of scores for a 48% shooter in 57 shots).

We can easily have the computer carry out this experiment many times, and keep track of the number of scores in each experiment:

```
bird = [48 1; 52 0];
z = starttally;
for expts = 1:1000  % do 1000 experiments
  a = sample(57, bird);
  baskets = count(a == 1);
  tally baskets z;
end
count( z<=20 )
```

We see that out of 1000 experiments the simulated normally playing Bird made 20 or fewer scores out of 57 shots about 40 times. This means that there is about a 4% chance of a normally playing Bird appearing to be in a “horrendous slump” when observed for 57 consecutive attempted shots.

What should we conclude from the data? Is Bird in a slump or not? All the above analysis tells us is that it is not impossible that Bird isn’t in a slump during the playoffs — he might be playing normally. [1]

So ... is Bird in a slump or not? Although one would like to give a precise yes-or-no answer, this is not really possible; we can answer questions only in terms of probabilities.

Statisticians have developed two types of framework for answering questions such as this one. By the word "framework" we mean a set of assumptions one makes and questions one asks about probabilities. Each framework is a kind of standard format for organizing information and providing quantitative answers to probability questions.

One type of question to ask about Bird's shooting data is this: "Given our observed data, what is the probability that Bird is in a slump?" This type of question is addressed by the *Bayesian* framework which we will explore in Chapter **??**.

For now, though, we will work in the *hypothesis testing* framework. This framework is motivated by a simple situation in logic. Suppose we have a hypothesis $\mathcal{H}$, for example, the hypothesis that "Bird is in a slump." If this hypothesis were true, then we expect a certain consequence: "Bird will shoot well below his normal 48% rate." This consequence itself implies something: "Our observed data will show that Bird scores only about 20 shots out of 57 attempts." We'll label the quoted statement $\mathcal{D}$ since it says what our $\mathcal{D}$*ata* will be if the hypothesis $\mathcal{H}$ is true. $\mathcal{D}$ is the *consequence* of our hypothesis $\mathcal{H}$. If $\mathcal{H}$ is true then $\mathcal{D}$ will be true. A shorter way of saying this is $\mathcal{H}$ implies $\mathcal{D}$ or, in symbolic form, $\mathcal{H} \to \mathcal{D}$.

Although the statement $\mathcal{H} \to \mathcal{D}$ may itself be perfectly true, this does not mean that $\mathcal{H}$ is necessarily true. It just means that if $\mathcal{H}$ were true then $\mathcal{D}$ would be true. In practice, we don't know whether $\mathcal{H}$ is true. We want to find out whether $\mathcal{H}$ is true by examining our data $\mathcal{D}$.

Our situation in statistical inference is that we observe some $\mathcal{D}$ and wish to infer something about the truth of $\mathcal{H}$. Students of logic have learned that $\mathcal{H} \to \mathcal{D}$ does not at all mean that $\mathcal{D} \to \mathcal{H}$. For example, the following statement is quite true: "$\mathcal{H}$: You are outdoors in winter in Minnesota implies $\mathcal{D}$: You will be cold." But, the fact that you are cold does not necessarily imply you are outdoors in Minnesota in winter. You might be in Alaska or even in Florida during a cold snap.

[1] Further reflection on this sitution suggests a related problem. Suppose you simulate a normally playing Bird for an entire season and look to see whether there occurs a sequence of 3 or more games where he is scoring very low, say below 30%. You will find out that it is almost certain that Bird will have such a sequence. This is good for reporters, who just by chance fluctuation will always end up occasionally being presented with a player in a "slump."

The statement $\mathcal{H} \rightarrow \mathcal{D}$ does however allow us to say one thing with certainty: not$\mathcal{D} \rightarrow$ not$\mathcal{H}$. If you are not cold then certainly you are not outdoors in winter in Minnesota.

This logical relationship between a hypothesis $\mathcal{H}$ and data $\mathcal{D}$ means that using data can in logic only refute a hypothesis. If the data are inconsistent with the hypothesis then we know that the hypothesis is wrong. If the data are not inconsistent with the hypothesis then ... well, the hypothesis might be right or wrong, we just don't know.

The hypothesis-testing framework has several components.

- The *Null Hypothesis* $\mathcal{H}_0$. This is a hypothesis that we will test. It should be something that it would be interesting to reject, typically a statement of the form "nothing is going on." We will check to see if the data are consistent with the Null Hypothesis. If they are not, we will conclude that the Null Hypothesis is false. "Bird is shooting normally" is an appropriate Null Hypothesis.

- The *test statistic*. This is the way that we summarize our data. In the basketball example, our test statistic is the fraction of scores our of 57 attempts. It is important to remember that the test statistic is not simply "the scoring rate" but is "the scoring rate out of a sample of size 57."

- The *p-value*. This is a probability. If the Null Hypothesis $\mathcal{H}_0$ is true, then there is a certain probability that, of all the data that we might have observed, any single data set would disagree with the Null Hypothesis to the same extent as the data we actually observed. In the basketball example, we found that by generating samples according to the Null Hypothesis that Bird is shooting at 48%, there is about a 4% probability that Bird would make 20 or fewer baskets out of 57 attempts. The p-value is therefore 4%.

  If the p-value is very low, one typically rejects the Null Hypothesis. The standard in scientific research is that the p-value should be below 5% in order to be justified in rejecting the Null. This is a valuable guideline, but there is nothing magic about the number 5%. If the p-value is above 5%, the Null Hypothesis may still possibly be false — it's just that our data doesn't justify this conclusion at the level scientists would typically find convincing.

Later, we will introduce some other terms relating to the hypothesis-testing framework: significance level, type I error, type II error, power. For the present, we illustrate some typical situations involving p-values and hypothesis testing.

3.2 Some Examples of Hypothesis Tests

Example 2: Testing a Difference in Proportions

"New polls show Ross Anderson surging in support in his campaign for President. Details at 11." This teaser for the 11 o'clock news does its job: you stay tuned to the television until 11, hoping for more information. If Anderson does stay in the campaign, he will split the political right and the centrists may win the election.

At 11:15 the details of the polls are given on the news show. Ten days before the election, Anderson had support of only 8%. Now, five days later, a new poll shows him at 11%. The news anchor reports that the polls were based on a random sample of 500 registered voters. Video footage is played of a spokesman for the Anderson campaign saying that the rapid growth in Anderson's support is due to increasing voter disenchantment with the big-party candidates as they face up to having actually to make a choice. The news anchor speculates that these "important" and "breaking" poll results show an "unprecedented growth in support for Anderson by almost 50%" and may herald a "new era in American politics." The next day, the Sunday talk shows are filled with discussion of this news. Even Bill Will, one of the most well respected commentators, says, "Something is definitely up. These polls are accurate to within plus or minus 2.5%, so the growth in support of 3% must be real."

You have a skeptical attitude. Nothing of significance or interest has happened in the campaign during the past week and voters are completely disinterested. You have a null hypothesis: the voters' level of support for Anderson hasn't changed in the past week. Are the data consistent with this hypothesis? You decide to check.

If your Null Hypothesis is correct, then the two polls are really two samples from the same population: nothing changed from the first poll to the second except that different random people were selected for the poll. Five-hundred randomly selected voters were polled each time, with 40 indicating support the first time and 55 the second time. Altogether, that makes 1000 people, of whom 95 supported Anderson: a 9.5% support rate. Taking this rate as indicating the voters' level of support for Anderson, you use resampling to simulate the polling process, and you look to see how often two polls will differ by 3 percent or more even though the underlying support rate stays constant at 9.5%. The file `pollchange.m` has the following lines:

```
voters = [9.5 1; 90.5 0];   % a 9.5 percent support rate.
Ntrials = 1000;
z = starttally;
for trials = 1:Ntrials
   % Conduct the first poll
  poll1 = sample(500, voters);
   % Now the second poll
  poll2 = sample(500, voters);
   % Compute the change in support level
  change = proportion(poll2 == 1)  - proportion(poll1==1);
  tally change z ;
end
disp('p-value of >= 3% growth in support:')
proportion(z >= 0.03)
```

Running the script

```
≫  pollchange  ⇒  ans:  0.052
```

The estimated p-value of the observed difference of 3 percent in Anderson's support is 0.052. This means that there is a reasonable chance (5.2% to be precise) that the second poll would show an increase of 3 percent or more in support compared to the first poll *even if the Null Hypothesis were true.*

What should you conclude? Since the p-value is not less than 5 percent, you would be justified in saying that the data do not cause you to reject the Null Hypothesis. But the computed p-value is awfully close to the conventional 5% cut-off for justifying the rejection of the Null. So you have little reason to insist that rejecting the Null is unreasonable. (Given the other information that you have — that nothing has changed in the campaign — it's probably sensible to wait for more information before concluding that there is a "new era in American politics.")

This situation of marginal p-values is quite common in hypothesis testing. Remember, that even if the Null Hypothesis is really true, one time in ten your data will produce a p-value less than 10 percent.

Many people are troubled by this situation, and want their data to lead to definite yes-or-no statements about the world. But it is frequently the case that the data are not so clear. One thing the marginal p-value does tell us in this case is that we really should have collected more data in the first place. We'll see an discussion of this in Example **??**.

Here is another thing to think about: If you take the point of view that the news interest in the polls would have been the same if Anderson's support had *fallen* by 3 percentage points, then the p-value should have been computed by counting any change whose absolute value was 3 percent or larger, that is:

```
≫   count( abs(z) >= 0.03 )
```

This revised computation gives a p-value of 0.11, which is not so marginal. The two cases, 1) considering an increase only, or 2) considering either an increase or a decrease, are called a *one-tailed test* and *two-tailed test* respectively. (One can also have a one-tailed test where only a decrease is considered.)

Which is better, a one-tailed or a two-tailed test? There is no hard-and-fast answer. The two-tailed test is more conservative, producing larger p-values. The one-tailed test, when appropriate, is more powerful. We'll return to this issue in Example **??** when we discuss the concepts of the *alternative hypothesis*, *power*, *Type I error* and *Type II error*.

Example 3: Labor Trouble: Difference in Two Means

Northworst Airlines is having labor difficulties. Six years ago they almost went bankrupt. Now, they are making huge profits and top management is giving itself multi-million dollar bonuses. Unfortunately, management didn't remember that these profits are due to low pay rates for mechanics, pilots, flight attendants and others who accepted deep pay cuts during the lean, almost-bankrupt years. The workers want a pay increase.

One way workers might be putting pressure on management is by causing delays — making flights reach the destination late so that customers become angry and switch airlines. Last month, a mechanics' union spokesman threatened management: either take negotiations seriously or the workers will start causing delays.

Northworst management has some data showing how late (in minutes) each flight is. To see if the union is making good on its threat, you have been asked to compare data from before the threat to data from last week. (These data and the analysis programs are in `airline.m`.)

```
beforethreat = [10 12 -1 82 7 -3 4 196 18];
afterthreat  = [-2 71 290 4 102 78 6 125];
```

(Negative numbers mean that the flight was early.) We can compare the mean delay before the threat and after.

```
≫   NWdiff = mean(afterthreat) - mean(beforethreat);
  ans:     48.1
```

Delays seem to have increased since the threat.

An appropriate null hypothesis in this case is that nothing has changed since the union threat was made, and that the before-threat data and the after-threat data are each random samples from the population of Northworst arrival times. If we assume that the null hypothesis of no change is correct, then our best guess about what the population of Northworst arrival times looks like is the combined before-threat and after-threat data.

```
≫   nullarrivals = concat(beforethreat,afterthreat);
```

One reasonable test statistic is the difference between the mean before-threat delay and the mean after-threat delay. As we saw, this value is 48.1 for the actual data. Now we want to assume that the null hypothesis is true and examine the distribution of values that the test statistic takes on. We can do this by resampling from `nullarrivals` and simulating a situation where a random sample of 9 arrival times is labeled as coming before the threat, and a random sample of 8 arrival times is labeled as coming after the threat.[2]

```
z = starttally;
for trials = 1:1000
   beforet = sample( 9, nullarrivals);
   aftert  = sample( 8, nullarrivals);
   teststat = mean( aftert) - mean(beforet);
   tally teststat z;
end
```

The array `z` now indicates the probability distribution of the test statistic *if the null hypothesis is true.* We want to see where the actual measured value of `NWdiff` falls in this distribution. This is shown in Fig. **??**, which was made with the following commands:

```
histogram(z,'Difference of means under Null Hyp.');
% Draw in the vertical line from 0 to 0.012, at 48.1
hold on;
plot( 48.1, 0:.001:.012, '*');
hold off;
```

[2] We have written the following MATLAB code using the numbers 9 and 8 to specify the sample sizes. This makes the code easier to read in this one example, but is bad programming practice in general. It would be better to write the two lines where sampling is done as:

```
beforet = sample( length(beforethreat), nullarrivals);
aftert = sample( length(afterthreat), nullarrivals);
```

```
xlabel('Diff in mean delays (minutes): after - before')
```

Figure 1:
Distribution of the test statistic under the null hypothesis that the before-threat and after-threat data come from the same distribution. The test statistic is the difference between the mean of 9 values minus the mean of 8 values. The value for the actual data is indicated by the starred vertical line at 48.1.

To compute the p-value, we want to see what fraction of the points in `z` are more extreme than `NWdiff`. In this case a one-tailed test seems appropriate since we have no reason to think the delay times would become shorter after the union threat. The one-tailed p-value is

```
≫   proportion(NWdiff<z)
  ans:     0.117
```

Just to illustrate how a two-tailed p-value would be computed in this case, we will go through the process involved in the calculation. First, we have to decide what we mean by “more extreme” when we say that we want to count the fraction of times that the null hypothesis produces a test statistic more extreme than those of the actual data. In this case, since `NWdiff` is larger than most of the values in the null's distribution, we'll take the right extreme to be `NWdiff`. (If `NWdiff` were smaller than most of the values in the null's distribution, we would take the left extreme to be `NWdiff`.)

For the extreme on the other side, we want to reflect `NWdiff` around the center of the distribution. `NWdiff` is displaced from the mean of the distribution by `(NWdiff - mean(z))`. Subtracting this from `mean(z)` gives us the left extreme. Translating this into MATLAB,

```
m = mean(z);
if( m < NWdiff )
   rightextreme = NWdiff;
   leftextreme = m - (NWdiff - m);
else
   leftextreme = NWdiff;
   rightextreme = m + (m - NWdiff);
end
```

and the fraction of points outside of the extreme values is

```
count( z< leftextreme | z > rightextreme)/length(z)
```

In this case, the two-tailed p-value is about 23%.

Since the p-value is so large even for the one-tailed case, we are not justified in rejecting the null hypothesis. Conclusion: there is no good reason to think that delays have increased since the threat. (But, even if we had found a small p-value, indicating that there was an increase that is statistically significant, we still wouldn't be able to be sure that the mechanics have *caused* the increase; all we can conclude is that there is an increase.)

It's not clear that the difference of means is the most appropriate test statistic to use in this case. Instead, we might want to use the difference in medians or perhaps the difference in the fraction of flights that are more than 15 minutes late. Historically, the emphasis on the difference in means stems from the fact that algebraic formulas are available for calculating p-values for the difference in means. There are no such formulas for most other test statistics. However, by using resampling it's straightforward to compute p-values for most any test statistic; just one line of the program needs to be changed:

```
teststatistic = mean(aftert) - mean(beforet);
```

could be changed to

```
teststatistic = median(aftert) - median(beforet);
```

or

```
teststatistic = proportion(aftert>15) - proportion(beforet>15);
```

or whatever statistic you prefer. Of course, you will have to use the same test statistic when computing `NWdiff` on the actual data. (And, when computing two-tailed p-values you should continue to use `mean(z)` to reflect around the center of the distribution.)

Which test statistic is most appropriate depends on what you think it is important to measure. If you think that really long delays are probably genuinely due to weather and other unavoidable factors, there is no reason for these to have a heavy influence on the results and the median delay is a better test statistic than the mean delay. If you think that union workers try to delay flights by 15-30 minutes so that connections to other flights are affected, then the fraction of delays more than 15 minutes might be the best test statistic.

There aren't very many data points in this record, and that can make the results of the calculations unreliable, particularly for statistics which count only a few of the data points. See Example ?? for an analysis of this.

CAUTION!

In the above example, we resampled *with replacement* from a Null Hypothesis data set that consisted of both the before-threat and after-threat data. For theoretical reasons, there is sometimes a small advantage in sampling *without replacement*. In the context of the above example, this is called a "permutation test." The idea is that we had altogether $9 + 8 = 17$ samples of data. Under the Null Hypothesis, any of these samples could have been collected before the union's threat, or after. To compute the distribution of the test statistic under the Null Hypothesis, we can randomly shuffle the 17 samples and pick 9 of them without replacement to call the before-threat "data." The remaining 8 samples are then the after-threat "data." Here's how to carry out the operation in the Resampling Stats software:

```
% for a permutation test
% construct the null hypothesis
nullarrivals = concat(beforethreat, afterthreat);
z = starttally;
for trials = 1:1000
   % randomize the order of the data points
   newdata = shuffle(nullarrivals);
   % pick out the first 9 of these to be the
   % simulated before-threat data
   beforeinds = 1:9;
   beforet = newdata(beforeinds);
   % take the remaining ones to be the
   % simulated after-threat data
   aftert  = exclude(newdata, beforeinds);
   % now do the test statistic, and so on...
   teststat = mean(aftert) - mean(beforet);
   tally teststat z;
end
```

Example 4: Testing a difference in paired data

Here is a very limited data set on the ages of husbands and wives:

```
husbands = [25 25 51 25 38 30 60 54 31 54 23 34 25 23
                 19 71 26 31 26 62 29 31 29 35];
wives =    [22 32 50 25 33 27 45 47 30 44 23 39 24 22
                 16 73 27 36 24 60 26 23 28 36];
```

The corresponding elements in the two arrays are from a married couple. For example, the first married couple consists of a man of age 25 and a woman of age 22. In the second couple the man is 25 and the woman is 32. (These data and the associated programs in this example can be found in `marriage.m`. If typing in the data by hand, type each set of data on a single line.)

A sociologist or an insurance company might be interested in examining such data to see if wives are systematically younger than their husbands. One way to do this (which we will see below is inappropriate for the husbands-and-wives data, but which is appropriate in many other cases) is to see whether the mean age of husbands is different from the mean age of wives. For these data

```
≫  agediff = mean(wives) - mean(husbands)
  ans:     -1.875
```

the wives are about 2 years younger than the husbands.

We want to know if this measured difference is statistically significant. That is, do these data provide evidence to refute the null hypothesis that husbands and wives do not differ systematically in age. If this null hypothesis is correct, then the `wives` and `husbands` data really come from the same distribution of data.[3] We can simulate this by combining the two data sets into one, and then resampling from this larger dataset. This forces our resamples to satisfy the null hypothesis.

First, we note the number of points in our data set:

```
≫  length(wives) ⇒ ans:  24
```

We'll concatenate the two data sets into one, and draw new samples from this combined data set.

```
% create the null hypothesis distribution
```

[3] Other null hypotheses can be imagined. For example, we might hypothesize that husbands and wives have the same mean age, but that the standard deviation is different for the two groups. In this case the data don't come from the same distribution; they come from two different distributions that have the same mean but different standard deviations. For ways of testing these hypotheses, see [?].

```
nullhypothages = concat(husbands, wives);
z = starttally;
for trials = 1:1000
   % we sample data from the null hypothesis distribution
  wifedata = sample( 24, nullhypothages);
  husbdata = sample( 24, nullhypothages);
  teststat = mean(wifedata) - mean(husbdata);
  tally teststat z;
end
% Compare z to the value of -1.875 found
% in the original data.
% Do a two-tailed test.
pvalue = proportion(z < -1.875) + proportion(z > 1.875)
```

We find a high p-value (0.625) indicating no systematic difference in the ages of husbands and wives.[4]

The above is a valid calculation which is appropriate for many data sets, wherever one wants to know whether two groups differ systematically from one another. In this case we use the mean as the test statistic, but the program is easily modified to work for other test statistics such as the median or standard deviation. Similarly, the two-tailed test could be changed to a one-tailed test by modifying the last line of the program.

Although the calculation does address the question of whether husbands as a group are different in age than their wives as a group, this is not the question in which we're interested. Instead, we want to know whether, *within each couple*, the wife is systematically different in age than the husband.

This is an example of a *paired* comparison. Paired comparisons are most frequently encountered in before-and-after experiments where a measurement is made on a subject, then some treatment is applied to that subject, then a second measurement is made. Since the same subject is involved in the two measurements, they are paired.

In our case the subject is a married couple. We are making a pair of measurements on each subject: the age of the wife and the age of the husband. Since we're interested in the age difference within a couple, we compute

[4] This example again violates good programming style by using the number 24 instead of `length(wifedata)` and `length(husbanddata)`. In addition, we've used the number `-1.875` instead of referring directly to the variable `agediff` which contains the value of the test statistic on the original data. We do this because the logic of setting left and right extremes for the two-tailed hypothesis test is a little bit complicated. See page **??** for an example of how this logic might be programmed.

```
≫   agediffs = wives - husbands;
```

As described in Step **??** in Appendix **??**, this commands computes the difference between the wife's and husband's age within each couple. Then, we can average this intra-couple difference over all of the couples.

```
≫   meandiff = mean(agediffs)
```

The result is -1.875, exactly the same as in the unpaired case! (If you remember about the "associative property" of addition and subtraction, and the "distributive property" of multiplication, it's easy to see why the difference of the means should be the mean of the differences.)

The important distinction between the paired and unpaired tests is not in the value of the test statistic, but in the manner in which statistical significance is calculated. When we did the unpaired calculation we set our null hypothesis to be that the age of each husband and each wife is randomly selected from a single distribution. According to this null hypothesis, a 73-year old wife is equally likely to be married to a 23-year old husband as to a 71-year old husband. This is obviously not the case.

For the paired calculation we respect the fact that the husband and wife in a married couple tend to be similar in age; our null hypothesis is that of the two ages for each couple, one is randomly assigned to the husband and the other to the wife. An easy way to implement this null hypothesis on the computer is to compute the age difference between wife and husband within each couple, and then resample by multiplying each age difference by either 1 or -1 randomly chosen. When multiplying by -1, we are effectively exchanging the ages of the wife and husband within each couple. Each resampled data set consists of all of the original data's age differences, but in a randomly selected set of couples the husband's and wife's ages are swapped.

```
z = starttally;
for trials = 1:1000
  signs = sample(length(agediffs), [-1 1]);
  teststat = mean( signs .* agediffs );
  % Note:  the . in .* means to use ordinary
  % element-by-element multiplication
  tally teststat z;
end
pvalue = proportion(z<= -1.875) +  proportion(z>rightextreme)
```

The p-value for the paired test is computed to be about 0.06 — which indicates a marginal degree of statistical significance for the finding that wives are about 2 years younger than their respective husbands.

3.2.1 Comparing two distributions

The next two examples show some approaches to exploring whether your observed data match a given probability distribution.

Example 5: Rolling a Die, revisited

Back in Example 1, we simulated rolling a die 6000 times. Since the outcomes of a single die are the numbers `[1 2 3 4 5 6]` with equal probability, each of the outcomes should appear roughly 1000 times. Of course, each value didn't appear *exactly* 1000 times because the sample is being drawn at random. The numbers we got in Example **??** were

Outcome	1	2	3	4	5	6
Observed Frequency	1014	958	986	995	1055	992
Expected Frequency	1000	1000	1000	1000	1000	1000

The question we want to answer here is this: are the observed results of the sampling so far from what is expected that the difference is too great to be accounted for by random sampling variability?

The question can be addressed by a hypothesis test in which the null hypothesis is that the observed frequencies do indeed match the expected frequencies and that any difference between the two is due solely to sampling variability. To carry out the hypothesis test we need a test statistic. The one we will use is the size of the difference between the expected frequency and the observed frequency, summed over all the outcomes. Since in some cases the observed frequency is higher than the expected frequency and in other cases it is lower, we'll use the absolute value of the difference to measure the size of the difference.

```
≫  observed = [1014; 958; 986; 995; 1055; 992];
≫  expected = [1000; 1000; 1000; 1000; 1000; 1000];
≫  result = sum( abs(observed - expected ))
  ans:     23
```

Now, the question is whether 23 is inconsistant with the null hypothesis. To answer this question, we'll generate many trials and check to see how often the size of the difference between the observed and expected frequencies is bigger than 23. The file DIETEST.M is an m-file script containing the following commands:

```
expected = [1000; 1000; 1000; 1000; 1000; 1000];
z = starttally;
for trials = 1:100
   rolls = sample(6000, [1 2 3 4 5 6]);
   observed = multiples(rolls);
   teststat = sum( abs( observed, expected ) );
   tally teststat z;
end

result = proportion(z > 23);
```

Executing this script, we get

```
≫  dietest ⇒ ans:  0.47
```

Approximately 50% of the time, the sampling variability produces a difference between observed and expected frequencies of more than 23. Thus, we have no reason to reject the null hypothesis.

If you have studied statistics previously, you may be familiar with the χ^2 ("chi-squared") statistic. This test statistic can be computed as

```
sum( ((observed - expected).^2)./expected )
```

Using χ^2 as a test statistic we again find that there is no reason to reject the null hypothesis. (χ^2 is traditionally the test statistic of choice in the type of situation in this example. When some of the expected frequencies are much smaller than others, χ^2 tends to weigh each of the outcomes equally, whereas the sum of absolute differences puts more weight on the more likely outcomes.)

Note that we're not testing the computer random number generator here, since the same random number generator that was used to create the original data is also creating the simulations we use for the hypothesis test. Instead, this example simply shows a technique for comparing a set of observed frequencies to some expected frequencies. But rest assured that the generator has been thoroughly tested, both by comparing it to other computer random number generators and by using tradition statistical tests such as χ^2.

Example 6: Grade Inflation

The legislature of a fictional eastern state has become concerned with grade inflation, and particularly the tendency for average grades at Eastern State University to increase each year at the same time as there is a perceived decrease in the skills of graduating students. To end this problem a new law has been passed, the "Eastern State Educational Standard Evaluation Statute," or ESESES, which stipulates in part:

> §17. Grades shall be assigned by each professor for each student based on the student's performance on accepted evaluative criteria without reference to the performance of the class as a whole.
>
> §18. The distribution of grades in each class shall conform to the standard distribution set out in §20 of this statute.
>
> §19. State University faculty whose class grade distribution substantially violates §18 of this statute shall be subject to administrative correction up to and including removal from office.
>
> §20. The standard distribution of grades in each class shall be the following:
>
> | A | B | C | D |
> |---|---|---|---|
> | 15% | 35% | 40% | 10% |
>
> [Ed. Note: This corresponds to a Grade Point Average (GPA) of 2.55.]

This is a difficult statute to comply with. On the one hand, §17 forbids grading on a "curve." (Grading on a "curve" means that a fixed fraction of students gets A, another fixed fraction gets B, and so on.) On the other hand, §18 says that in the end the grades awarded must somehow align themselves perfectly with the curve specified in §20.

The state Attorney General, when prosecuting faculty, has defined a "substantial violation" to be a difference of 0.5 in the class' mean GPA from the legislatively prescribed class GPA of 2.55. This means that any professor teaching a class the average of which is above 3.05 or below 2.05 is subject to being fired.

A young statistics professor has received an administrative complaint. The grades in her early-morning class of 10 students were 3 Cs, 3 Bs and 4 As, giving a class GPA of 3.10. In responding to the complaint she wrote:

"The ESESES statute needs to be understood in a statistical context. Insofar as students are randomly assigned to a class, and insofar as the professor assigns each grade independently as required in Section 17, the class' GPA will be a random variable. By chance fluctuation, the GPA could be *substantially* different than the legislature's target of 2.55, but possible not be *significantly* different. To illustrate I have used the Resampling Stats software to simulate the case where students are drawn randomly from a population that exactly matches the legislative standard. [The program is in gpadist.m.]

```
classsize = 10;
% the legislatively mandated distribution of grades
standard = [.15 4; .35 3; .40 2; .10 1];
z = starttally;
Ntrials = 1000;
for trials = 1:Ntrials
   class = sample(classsize, standard);
   teststat = mean(class);
   tally teststat z;
end
res = proportion( z > 3.05 | z < 2.05);
```

"For my classsize of 10 students, this calculation indicates that in 8% of classes where the individual students' grades come from a population that complies exactly with the mandated standard there will nonetheless be an apparent violation of the Attorney General's standards. Another, similar calculation shows that the two-tailed p-value for my class' mean GPA of 3.10 is only 5.5% and therefore within the statistically acceptable bounds for compliance.

"Note that the Attorney General's criterion relates to the 'substantiality' of a GPA difference: whether it is big enough to be of interest. My use of the word 'significance' in its statistically correct meaning relates to whether the measured difference is big enough to be reliably discernable from no difference at all. Given the size of my class, the difference in GPA from the mandated value is too small to conclude that I have violated the legislative standards."

To analyze the professor's case it is worthwhile to point out the parallels between the hypothesis-testing framework and the well known standards for trial in American courtrooms:

- "Innocent until proven guilty." The professor is saying that the null hypothesis should be that she has complied with the legislatively prescribed grade distribution. Only if the data strongly indicate that the null hypothesis should be rejected — that is, if the p-value is very small — should she be found guilty.

- Guilt must be proved "beyond a reasonable doubt." Even if the null hypothesis is true (that is, the professor is innocent), there is some probability of the observed evidence. This probability is the p-value. In order to reject the null hypothesis of innocence, the p-value must be very small — beyond a reasonable doubt.

Note the strategy used by the professor in computing the p-value. She had two sets of numbers: one is the grade distribution specified by the state legislature, the other is the 10 grades in her own class. To compare these two sets, she reduced each to a single value, the mean of the set. This is appropriate in this case because the Attorney General defined compliance in terms of mean GPA.

More generally, one might like to know whether the observed distribution of grades in the class corresponds to the distribution mandated by the state legislature. This problem is identical in spirit to the one examined in Example **??**, where we compared the frequencies of the actual outcomes of die rolls to those expected theoretically. In the case of the grade standards, requiring that the distribution of grades match the legislative standard is more restrictive than requiring only that the class GPA match — not merely must the mean grade meet the standard, but also that there be the required number of A's, of B's, of C's and of D's. But, since in §17 of the statue it says that the professor cannot simply have a grade quota — each student must be graded individually. This means that there will be some fluctuations in the number of A's, B's, C's and D's, depending on which particular students happened to be in "random sample" that composes the class. So, instead of giving a yes-or-no answer to the question of whether the class' grades match the standard, we need to compute a test statistic that quantifies the mismatch, and calculate the p-value of this test statistic under the null hypothesis that the class' grades do indeed come from a population matching the legislative standards. As discussed in Example **??** two relevant test statistics are the sum of absolute deviations from the expected number of grades of each type, and the χ^2 statistic. The following program calculates the sum of absolute deviations:

```
classsize = 10;
% The legislated distribution of grades
standard = [.10 1; .40 2; .35 3; .15 4];
```

```
% This give the expected number of each letter grade
% in the same order as standard
expectednum = classsize*[.10 .40 .35 .15];
z = starttally;
for trials = 1:1000
   % generate a simulated class
   simclass = sample(classsize, standard);
   % compute the observed number of grades of
   % each type
   observednum = multiples(simclass, [1 2 3 4]);
   teststat = sum( abs( observednum - expectednum ) );
   tally teststat z;
end
```

(The above script is found in gpa2.m.[5])

After running the script, the tallying variable z contains a sample from the distribution of the test statistic under the null hypothesis. The value of the test statistic for the actual class data (0 D's, 3 C's, 3 B's, 4 A's) is

```
≫  sum(abs([0; 3; 3; 4] - expectednum))
  ans:     5
```

The one-tailed p-value is

```
≫  proportion(z >= 5)
  ans:     .402
```

The conclusion: there's no reason to believe that the professor's grades are not a random sample from the distribution specified by the state legislature.

When using test statistics like the sum of absolute deviations or χ^2, the test is typically one-tailed, since any deviation from the expected numbers will lead to a positive value for the test statistic. The larger the value of the test statistic, the more evidence that the observed distribution is not the same as the expected distribution. There is one case, however, when you might be interested in very small values of the test statistic; when you suspect that the observed and expected distributions match too closely. This can be an issue, for example, when investigating scientific fraud and the fabrication of data.

[5] It is better style to write the program a bit differently. One should write

```
expectednum = classsize*standard(:,1);
```

which uses the first column of standard to compute the expected number of grades in each type. Similarly, one should write

```
observednum = multiples(simclass, standard(:,2));
```

This avoids mistakes arising from typing the same information in two different places.

All of the calculations here are based on the assumption that the students in the class represent a random sample of all students. This is unlikely to be the case. If a subject is difficult, then better students will tend to take the classes in that subject. If a class is offered in the early morning, then more motivated students tend to take the class. Just for these reasons, the ESESES statute is flawed. But this doesn't mean that the calculations shown here are useless: the sampling variability puts a lower bound on how much the grade distribution can be expected to differ from the legislative standard. If the actual deviation is larger than the deviation expected from sampling variability, we then need to move on the the issue of what caused the larger-than-expected deviation.

CAUTION!

For example, suppose there is a class at the State University with 100 students, and that the grade distribution was 149 A's, 351 B's, 400 C's and 100 D's. This matches the legislative standard almost exactly, but perhaps it is evidence that §17 of the statute is being violated. The sum of absolute deviations is easy to calculate: there is one too few A's and one too many B's, so the sum of absolute deviations is 2. Running the above script again, but setting `classsize` to 100, we get a sample from the distribution of the test statistic under the null hypothesis. We compare this to the observed value of 2 for the test statistic:

```
≫  proportion(z <= 2)
  ans:     0.016
```

The low p-value indicates that the recorded grades are suspiciously close to the standards, and suggests (but does not prove absolutely) that there was a grade quota that was adjusted so that the numbers didn't exactly match the standard.

We certainly do not advocate cheating and falsifying data, but you should be aware that a clever cheater will adjust things so that a test like this won't give evidence for cheating. In the case here, a reasonable amount of adjustment is

```
≫  median(z)
  ans:     12
```

So, a clever cheater will arrange things so that the sum of absolute deviations is around 12.

3.2.2 *Independent resampling of two variables*

When examining the relationship between two variables, we typically make a measurement of both variables for each case. A typical null hypothesis is that there is no relationship between the two variables. In order to generate resampled data that is consistent with this null hypothesis, we can resample separately from each of the variables. This is shown in the next example.

Example 7: Testing a correlation

We revisit the stork and baby data. In Example **??** we were concerned with the confidence intervals on a statistic, r^2, that describes how well the data are modeled by a straight line. Here, we'll consider the situation from the perspective of hypothesis testing. Our null hypothesis is that there is no relationship between births and the stork population. As a test statistic we will again use r^2, although we could use other statistics such as the slope of the line that best fits the data.

To implement the null hypothesis, we resample independently from the stork data and from the baby data. Doing so means that our resampled data avoids any real correlation between the two data sets and allows us to test whether the observed degree of correlation — $r^2 = 0.89$ as found in Example **??** — could occur by chance. The program `storkscorr.m` is almost exactly the same is in Example **??**, but the three lines in the original program that randomly selected pairs of data points

```
inds = sample( Ndata, 1:Ndata );
stks = storks(inds);
babs = babies(inds);
```

are replaced with independent samples from the stork data and births data separately.

```
stks = sample( Ndata, storks );
babs = sample( Ndata, babies);
```

A one-tailed test is appropriate here, since large values of r^2 indicate that there is a correlation, but small values of r^2 do not. The p-value comes out to be about 1%, indicating that the null hypothesis can be rejected. This is essentially the same conclusion we reached from the confidence interval calculation in Example **??**.

3.2.3 Adjusting for multiple tests

In many studies, one has many different variables available, and perhaps many different test statistics. It is tempting to search through the data, looking for those variables that show something interesting. This is a legitimate method of data exploration, but great care must be taken when computing and interpreting the p-values in such situations.

Example 8: Discounting Multiple Comparisons

You have just read in your fictional local newpaper a fascinating study that shows that boys are different from girls in their performance on tests of spatio-temporal reasoning. In the study, scientists took a randomly selected group of 3000 children, half boys and half girls, and subjected them to a battery of 15 tests. The score on each test was the time taken to perform a task involving spatio-temporal reasoning. Some of the tasks were: assembling a small jigsaw puzzle; sorting differently colored marbles according to size; building a specified shape out of blocks, and so on. In both the jigsaw puzzle task and the marble-sorting task, girls were faster than boys. The one-tailed p-values for these two tests were 0.03 and 0.04 respectively. The other tasks showed interesting patterns also, but none that reached statistical significance at a level $<$ 0.05. In particular, in none of the 15 tests did the boys perform faster than the girls at a statistically significant level.

The newspaper touts this study as "scientifically rigorous and convincing with a high level of statistical significance." Certainly p-values of 0.03 and 0.04 seem to be compelling. It's also interesting that both tests point in the same direction: girls being faster than boys. One of the scientists in the study is quoted as saying, "The fact that two tests separately give highly significant results, means that the overall significance of the study is wildly significant, at a level of $p < 0.0012$."

You are not so sure. If 15 tests were done, it doesn't seem quite right to focus on those 2 tests that gave the strongest results. With 15 tests, it seems that there is a good possibility that at least one of the tests would prove significant even if there is no real difference in spatio-temporal reasoning. But is it likely that two tests would be significant and in the same direction? What about the fact that none of the tests were significant in favor of boys' being faster?

Let's answer these questions with a simulation. We note that if the null hypothesis is true, the one-tailed p-value is a random variable uniformly distributed between 0 and 1. (Did you think that the p-value will tend to be close to 1 if the null hypothesis is true? Not at all.)

Here is a thought experiment to show that the above statement is true. Imagine that there is a hidden switch that controls whether the null hypothesis is true or false. We carry out a study by collecting data and computing single test statistic, for instance the mean difference in time to complete the task for boys versus girls. We'll call this test statistic the "observed result."

Now, set the switch to TRUE and collect another 100 values of the test statistic. (In the previous examples in this book, we have done this by simulation and resampling.) The 100 values will vary randomly from one another, but they all reflect the distribution of the test statistic under the null hypothesis. We will call the 100 numbers the "null distribution."

The one-sided p-value of your observed result can be estimated by counting how many of the numbers in the null distribution are smaller than our observed result. If none of them are smaller than the observed result, the p-value estimate is 0.00. If 1 of them is smaller, the p-value estimate is 0.01, and so on. If all of them are smaller, the p-value estimate is 1.00.

Suppose now that the switch had been set to TRUE when collecting your observed result. Then all 101 values of the test statistic are drawn from the same probability distribution. The situation is analogous to putting 101 different numbers in a hat, and drawing one of the numbers at random to be your observed result. The probability that the drawn number is the smallest of all is $\frac{1}{101}$. The probability that the drawn number is the next-to-smallest is $\frac{1}{101}$, and so on. So, if the null hypothesis is true, each of the possible p-values from 0 to 1 is equally likely.

In our simulation, we will generate 15 p-values uniformly distributed between 0 and 1. We'll find the probability tha there is one value less than 0.03 and one less than 0.04, and that none of the 15 p-values are greater than 0.95. (A p-value of 0.95 from a one-sided test that girls are faster than boys corresponds to a 0.05 p-value that boys are faster than girls.)

Here is the program in `manytest.m`:

```
z = starttally;
for trials = 1:1000
   pvalues = uniform(15,0,1);
   % sort from smallest to largest
   pvalues = sort(pvalues);
   goodenough=pvalues(1)<.03&pvalues(2)<.04 & pvalues(15)<.95;
   tally goodenough z;
end
```

```
proportion(z)
```

The only new programming technique here involves sorting. We sorted the randomly generated p-values from smallest to largest so that we can easily find the smallest and second smallest p-values to compare them to 0.03 and 0.04. We also check whether the largest p-value is less than 0.95; if this is true then all of the p-values are less than 0.95.

The result: the p-value for the overall reported results of the study is 0.06. This is not as significant as the 0.03 reported for the best test and quite different than the 0.0012 reported. If you take the point of view that all of the tests should have been done two-sided, the overall p-value should be $2 \times 0.06 = 0.12$ which is not very compelling evidence for a difference between boys and girls. The two-sided is computation is appropriate, for instance, if you don't know ahead of time that girls will be faster than boys, but would be equally willing to consider significant a finding that boys are faster than girls.

When embarking on a long study, keep track of how many hypotheses have been tested and make your criterion for rejection of the null hypothesis in each case more rigorous. A simple, but perhaps overly conservative way of doing this is the Bonferroni correction: reduce the significance level by an amount proportional to the number of hypotheses tested. For example, if your study will investigate 10 hypotheses, and a 5% significance level is desired, then use a 0.5% level in conducting the individual tests. A discussion of these issues is contained in many statistics books, e.g., [**?**, **?**].

Another approach is this: When you have found a hypothesis that appears worthwhile, go back and collect new data for conducting the hypothesis test. In many cases, researchers divide their original data set into two parts. One part is called the *training set* and is used for exploring many different hypotheses. The other part is called the *testing set* and is used for conducting a new hypothesis test on those few hypotheses that seemed worthwhile based on the analysis of the training set.

Be wary of research reports where multiple hypotheses have been tested and no explicit adjustment has been made to the significance level.

CAUTION!

3.2.4 Sample size and power

The next two examples deal with an important issue in designing studies or experiments: how many data samples should we collect if we want our study to be a success. By "being a success" we do not mean exactly that the study will prove the outcome we want. Honest studies are designed so that they reflect the world as it actually is and not just the way we would like it to be. Instead, "being a success" means that our result should fail to reject the null hypothesis if the null hypothesis is true, and should lead us to reject the null hypothesis if it is false.

We have already seen how to use p-values to guard against falsely rejecting the null hypothesis. The common technique of the previous several examples was to compare the observed value of our test statisic from the real data to the test statistic from many trials of simulated data generated in a way consistent with our null hypothesis. We use this comparison to generate a p-value that describes how likely is our observed value of the test statistic if the null hypothesis were true. A very low p-value indicates that the data are inconsistent with the null hypothesis.

Although p-values help to solve the problem of false rejections a true null hypothesis, we also face a different problem: we don't want to fail to reject the null hypothesis if it is false. How do we design studies so that they will reject the null hypothesis if it is false? The general strategy is this: generate simulated data that is consistent with the null hypothesis' being *false*. Then compute the p-value of this simulated data.[6] We carry out many such trials, computing a p-value for each of many simulated data sets. We want the study design — specifically the number of data samples in our study — to be such that for our simulated data there is a large probability that the p-value we compute from the simulated data will be low enough to reject the null hypothesis.

Unfortunately, saying simply, "the null hypothesis is false," does not tell us exactly what is true. In order to do calculations, we need a specific statement about what is true. This specific statement is called our *alternative hypothesis*.

If the real world is like our simulations of the alternative hypothesis, then our simulations should show the probability that our real-world study will lead us to reject the null hypothesis. This probability is called the *power* of the study. We want to set our study's sample size to be big enough to make the power close to 100%.

[6] There is a twist here: the p-value computation is done by simulating data that is consistent with the null hypothesis' being *true*.

Example 9: Designing a Medical Study

You work for a large pharmaceutical company that has developed a new drug to combat hypertension (high blood pressure). Preliminary studies have shown that the drug is acceptably safe in humans. Your job is to design the study that will convince the government's Food and Drug Administration (FDA) that the drug is effective.

You are planning a double-blind, controlled study. In this study, volunteer patients with high blood pressure will be randomly assigned either to take the new drug or a placebo. Neither the patient nor the patient's physician will know whether the patient is taking the drug or the placebo. At the beginning of the study, each patient's blood pressure will be measured. After two weeks on the drug, the patient's blood pressure will again be measured and the change in blood pressure will be recorded. These data will be sent to you for analysis. Only you know which patients received the drug and which received the placebo.

The question you face is how many patients to enroll in the study. Because of the risk of side effects from the drug, you want to keep the study as small as possible. On the other hand, you don't want the study to be so small that the results are not statistically significant.

A preliminary study, with 10 subjects, showed that the drug reduces systolic blood pressure by approximately 14 mmHg.[7] Unfortunately, the p-value for that study was .15; not low enough to establish statistical significance. Let's assume that the FDA requires a p-value of 0.025 for a one-tailed test. The FDA presumes that you know in which direction the drug will act, hence a one-tailed test is appropriate.

From your perusal of the medical literature, you know that in the hypertensive population from whom the drug is intended, measurements of blood pressure made two weeks apart — with no treatment — differ by 0 mmHg on average with a standard deviation of 21 mmHg. You also know, from previous studies, that the placebo effect applies, and that giving a placebo will reduce blood pressure by 4 mmHg, on average. We will use these facts to constitute our *alternative hypothesis.*

You can approach the problem using a simulation. Here is a program (found in file `bpstudy.m`) that will generate data for `N` subjects, and compute the significance level of the difference between the placebo and control groups. The data are generated to be consistent with the alternative hypothesis.

[7] Blood pressure is measured in millimeters of mercury (mmHg). Systolic pressure is the peak blood pressure after each heart beat.

```
function pvalue = bpstudy(N)
% bpstudy(N)
% simulate the blood pressure study with
% N subjects receiving the drug and
% N subjects receiving the placebo.
% the returned value is the p-value of the
% difference.

% Simulates the before-and-after difference in blood pressure
% generate data for the placebo group
placebo = normal(N, -4, 21);
% and for the drug group
drug = normal(N, -14, 21);
% Now compute the p-value of the diff. between the two groups
% Use a permutation test
observedval = mean(drug) - mean(placebo));
z = starttally;
nulldata = concat(placebo, drug);
for trials = 1:500
  pinds = 1:length(placebo);
  newdata = shuffle(nulldata);
  sampplacebo = newdata(pinds);
  sampdrug = exclude(newdata, pinds);
  teststat = mean(sampdrug) - mean(sampplacebo);
  tally teststat z;
end
% The observed difference in means is expected to be < 0.
% So we'll look for trials where z is even more negative.
pvalue = proportion(z < observedval);
```

Let's try it out with a sample size of 10:

```
≫  bpstudy(10)  ⇒  ans:  0.5520
```

The result is somewhat random, so let's try again:

```
≫  bpstudy(10)  ⇒  ans:  0.1280
```

The second simulation is quite different from the first. This isn't a problem with the number of trials in the permutation method. Instead, it genuinely reflects the fact that the small population in the study makes it hard to get a significant result.

We'll try a larger population:

```
≫  bpstudy(50)  ⇒  ans:  0.044
≫  bpstudy(50)  ⇒  ans:  0
```

It looks like a study involving 50 subjects will give a low p-value but we're not confident that the p-value would be lower than the FDA-mandated 0.025. To make sure, we'll write a program to carry out many such simulations for a given size. The following program is in file `bppower.m`:

```
function res = bppower(studysize)

% carry out many simulations
z = starttally;
for trials = 1:100
   pvalue = bpstudy(studysize);
   tally pvalue z;
   pvalue     % print out intermediate results
end           % because the program is quite slow
% count what fraction of studies gave
% a sufficiently low p-value
res = proportion(z <= 0.025);
```

Each time we run BPPOWER, it carries out 100 simulations of our experiment and reports the fraction of those experiments that gave a sufficiently large p-value to lead us to reject the null hypothesis. Let's try a study size of 30 subjects (in each of the two groups):

```
≫ bppower(30) ⇒ ans: 0.32
```

We see that if we run a study with 30 subjects in each group, there is about a $\frac{1}{3}$ probability that the study will be a success. By "success" we mean this: if the alternative hypothesis is indeed true, then the study is a success if it leads us to reject the null hypothesis. The probability that the study will be a success in this sense is called the *power* of the study design.

A power of $\frac{1}{3}$ might be adequate for some purposes, but one wants the power to be as close to 1 as possible (consistent with real-world constraints such as the cost of recruiting additional subjects, and so on). Let's try a study size of 50:

```
≫ bppower(50) ⇒ ans: 0.62
```

We see that increasing the number of subjects increases the power. Let's make it bigger still:

```
≫ bppower(70) ⇒ ans: 0.78
```

It seems that with 70 subjects in each group, the power of the study is roughly 78%. This seems pretty large, but keep in mind that it means that even if our alternative hypothesis is in fact correct, there is a 22% chance that our study (with 70 subjects) will fail to reject the null. We'll try a much larger study:

```
≫ bppower(250) ⇒ ans: 1.00
```

By trying many different sample sizes, we can find one that gives us the power we want with the budget we have. (Note: Don't take that 1.00 exactly literally: it just means that in our 100 trials we didn't encounter even a single case of the simulated study failing to reject the null hypothesis. We need to increase the number of studies to have more precision in our estimate of the power. But remember that this estimate is based on the assumptions that went into the simulation; these may or may not be accurate.)

Example 10: The Statistical Power of Polling

The Presidential election has come and gone. Anderson got 9% of the vote. He didn't qualify for federal campaign support in the next election, but he did split the right-wing vote.

On Public Television, Bill Moyers is hosting a round-table discussion about the media coverage of the campaign. The issue is Anderson's reported "surge" in support in the last days of the campaign. (See Example **??**.) As facts turned out, the surge never materialized. Reporters have egg on their faces.

A statistician is critical of the reporters: "There was never any reason to report a surge. The p-value of the reported surge was 11 percent — no reason to reject the Null Hypothesis of no change in support."

Cynthia Brokow retorts, "That's for a two-sided test. The one-sided p-value was about 5 percent. It's our responsibility to keep the public informed and not to suppress information because it doesn't reach some ivory-tower threshold for reliability. We gave the raw numbers from the poll; it's up to the viewer to figure this out."

Moyers mediates. "Clearly there's a problem here. We reported a story that was wrong and for which, in hindsight, we didn't have much evidence. If we reporters can't digest these statistics, how can we expect the public to do so? There must be some balance between reporting the raw facts and reporting only those facts which, with due statistical consideration, provide a reasonable level of support for the conclusions they seem to point to."

Steven Brill, editor of a media watchdog magazine, has a suggestion. "This is a question of standards and responsible reporting. We have an obligation to collect enough data to make our results reliable, particularly when the results are important. The problem is in the size of the poll.

The polls have to be big enough so that we when claim something as remarkable as a 3 percent increase in support we have good reason to believe the data. I don't care if we make mistakes with claims concerning 1 percent changes in support, but we have to be right when claiming 3 percent."

Moyers: "Well, how big does such a poll have to be?" All eyes turn toward the statistician.

The statistician: "This is an example of a *sample-size* calculation. We want to make the sample size large enough so that our hypothesis test has a low *significance level* against the null hypothesis and a high *power* against the alternative hypothesis. As you know, there's generally a trade-off between power and significance, and ..."

Moyers interrupts. "Hold on a second. Let's bring this down to Earth. I don't know much about statistics but as a reporter I know that we want our stories to have a high significance level."

The statistician: "Sorry. I was using some technical terms whose meaning doesn't always correspond well to the everyday meaning of these words. The *null hypothesis* is a statement which we are going to *reject* or *not reject* on the basis of our data."

Reporter: "Like, 'Nothing much has happened. No change in support."

Statistician: "Exactly. The null hypothesis plays the role of the devil's advocate. We also have a *test statistic* — in this case that's the fraction of support measured in our poll, or, rather, the change in the fraction of support between the two polls. And, we have a *rejection threshold* that measures what we're interested in. This is a level we set ahead of time. If our test statistic is beyond the threshold level then we conclude that the data justifies rejection of the null hypothesis."

Reporter: "What about the p-value?"

Statistician: "The p-value is something we calculate after we already have our data. Right now we're discussing how to design the poll, not how to analyze the data from the poll."

Moyers: "So where does the sample size come in?"

Statistician: "The sample size determines where we set our rejection threshold so that our conclusions are reliable. Imagine that the devil's advocate is right and the the null hypothesis is true. Since we're randomly picking the voters questioned in the poll, it might happen that our poll results are above the rejection threshold just by chance. So, even if the null hypothesis is right our poll results might cause us to reject the null."

Sam Donaldson: "That's pretty unlikely."

Statistician: “I don’t know how you can say that since we haven’t yet set either the threshold or the sample size. Of course, you’re right in the sense that our goal here is to set the sample size and threshold so that the probability of falsely rejecting the null hypothesis is very small. A false rejection of the null is called a *Type I error* and the probability of making such an error — for a given null hypothesis, threshold, test statistic, and sample size — is called the *significance level* of the test. We want a low significance level, that is, a low chance of making a Type I error.”

Moyers: “How do we find the significance level?”

Statistician: “Since the sample size is what we want to figure out, first we pick a rejection threshold.”

Moyers: “OK. How do we pick a rejection threshold?”

Statistician: “Bear with me — we’ll come to that in a bit. For now, let’s assume that you already have the threshold, say a change in the polls of 1%. This means that if the test statistic — which is the difference between successive polls — is more than 1%, we will reject the null hypothesis. We want to make the sample size large enough to make the significance level small.”

Moyers: “How small a significance level is small enough?”

Reporter from the *McLaughlin Group*: “50%”

Reporter from CBS news: “25%. We have high standards.”

Reporter from the *Christian Science Monitor*: “0%. We want to be right all the time.”

Statistician: “The standard in scientific research is 5%. This means that if the null hypothesis is true, we’ll make a mistake about one time in 20. If we set the significance level at 50%, even when nothing is happening we will have a headline story — albeit wrong — for just about every second poll.”

National Inquirer: “Exactly. That’s the nature of our business. The public has a right to know.”

Statistician: “On the other hand, it would be a mistake to insist that we never make a Type I error for example insisting on a 0% significance level. Doing so would practically ensure that we would always make Type II errors.”

Moyers: “Type II errors?”

Statistician: “A *Type II error* occurs when the null hypothesis is wrong, but we fail to reject the null.”

Moyers: “That would happen, for instance, if there were a big change in support from one poll to the next but our rejection criteria were so rigorous that we refused to conclude that something had changed.”

Statistician: “Right. What we want to do is set our rejection threshold to make both types of error unlikely. Unfortunately, there is a trade-off between making the two types of error. For instance, we can lower the probability of a Type I error by making the rejection threshold harder to satisfy.”

Moyers: “You mean by saying that we won’t report that there has been a change in support unless the difference from one poll to the next is at least 2%, not 1% as previously suggested.”

CBS news: “But that would make it less likely that we’d be able to report a change.”

Statistician: “Right. That would be good, though, if there really were no change. You’d avoid a Type I error.”

CBS news: “But what if there were really a change in support?”

Statistician: “Then not reporting it would be a Type II error. As I said, there’s a trade-off between Type I and Type II errors. If you alter the rejection threshold to reduce the probability of making one type of error, you increase the probability of the other type.”

Moyers: “Fascinating. But where does the sample size come in?”

Statistician: “There is one way around the trade-off. We can reduce the probabilities of both types of error by making the sample size large.”

Moyers: “How large?”

Statistician: “The larger the better. But in order to make the poll economically feasible, you also want to make the sample size small. So, I’ll calculate the minimum acceptable size of the sample. First, I need to compute the probability of a Type I error. What’s your null hypothesis?”

Sam Donaldson: “That there has been no actual change in the level of support.”

Nina Totenberg: “But what will we be justified in reporting if we reject the null; only that ‘support has increased.’ That’s not a very strong statement.”

Statistician: “Right. Perhaps you’d rather have a stronger statement. If your null were ‘support has changed by less than 1%’ then if you reject the null you’ll be able to make a stronger statement.”

Moyers: “Let’s take ‘less than 1%’ as our null.”

Statistician: “We’ll use a significance level of 10% for the calculation. Now ... What’s your alternative hypothesis?”

Moyers: “You mentioned that at the beginning. What is that?”

Statistician: “The *alternative hypothesis* is something that, if true, would lead you to reject the null.”

Moyers: “Why not just take the alternative to be that the change in political support was greater than 1%. That’s what we know if the null isn’t true.”

Statistician: “Good point. However, I need a specific hypothesis so that I can calculate the probability of a Type II error. Is it alright if I say that the alternative is, ’The real change in support was 3%?’ ”

Tottenberg: “Why not say 2.1%?”

Statistician: “We could. But before deciding, let’s pick an acceptable error rate for Type II errors. If there really was a change of 3%, how much chance are you willing to take that you make a mistake and fail to reject the null?”

Moyers: “That’s difficult to answer. Failing to report something doesn’t seem like as serious an error as reporting something that is wrong. Let’s say that we’re willing to miss the story 25% of the time.”

Statistician: “OK. I have the information I need. By the way, the *power* of the hypothesis test is 1 minus the probability of a Type II error. That’s 75% in this case and is the probability that we do (correctly) reject the null when the alternative is true.”

“I’ll do the calculations for a case that’s like the Ross Anderson situation where the background level of support is about 10%.”

The statistician writes a program in Resampling Stats. It can be found in `pollsize.m`.

```
function [thresh, type2rate] = pollsize(samplesize)
% Find the rejection threshold and Type II error rate
% of a test statistic which is the difference between
% successive polls.
backgroundsupport = 0.10;
nullincrease = 0.01;
alternativeincrease = 0.03;
significancelevel = 0.10;

z = starttally;
Ntrials = 1000;
for trials = 1:Ntrials
   % simulate the difference of two polls under the null.
   poll1 = sample(samplesize, ...
      [backgroundsupport 1; (1-backgroundsupport) 0]);
   poll2rate = backgroundsupport + nullincrease;
   poll2 = sample(samplesize, [poll2rate 1; (1-poll2rate) 0]);
   teststat = (count(poll2==1) - count(poll1==1))/samplesize;
   tally teststat z;
end
% compute the rejection threshold you’d have to use to make a
% mistake at the specified significance level
% since we are testing only for an increase in support, we look
% at the right side of the distribution of z
thresh = percentile(z,1-significancelevel);

% Using the threshold, compute the Type II error rate under the
% alternative hypothesis.
z2 = starttally;
for trials = 1:Ntrials
   poll1 = sample(samplesize, [backgroundsupport 1; (1-backgroundsupport) 0]);
   poll2rate = backgroundsupport + alternativeincrease;
   poll2 = sample(samplesize, [poll2rate 1; (1-poll2rate) 0]);
   teststat = (count(poll2==1) - count(poll1==1))/samplesize;
   tally teststat z2;
end
type2rate = count(z2<thresh)/length(z2);
```

This program will take any sample size and compute the rejection threshold and the Type II error rate. It's assumed that the significance level is 10%. We try this out for many sample sizes until we find the smallest one that gives us a reasonable Type II error rate. Then we just read off the appropriate rejection threshold.

Statistician: "Let's try a sample size of 100 (in each poll).

```
≫  [thresh, errorrate] = pollsize(100)
  ans:     0.06 0.70
```

We get a threshold of 6% and a Type II error rate of 70%. "

Sam Donaldson: "You mean that we won't say that there has been a change in support unless the polls have changed by 6%. That's ridiculous."

Statistician: "I agree. It means that the sample size is too small. Let's try 500.

```
≫  [thresh, errorrate] = pollsize(500)
  ans:     0.034 0.570
```

Now the threshold is 3.4% and the Type II error rate is 57%. This is still much too high. So let's try a much larger sample size.

```
≫  [thresh, errorrate] = pollsize(2000)
  ans:     0.023 0.230
```

Good. The Type II error rate is down to 23%, close to the specified value. The threshold is 2.3%. That seems to fit the bill, but barely. So, this is the smallest sample size that's acceptable."

Moyers: "I notice that you didn't really use our null hypothesis that support changed by less than 1%. Instead, you assumed that support changed by exactly 1%. Why?"

Statistician: "I wanted to make the computations conservative, so I took the worst possible case."

Moyers: "But how do you know this isn't too conservative. It's awfully expensive to poll 2000 people."

Statistician: "In order to do the calculation differently, I would need some more information: under your null hypothesis how likely is it that the real change is 0%, 1%, and so on. I don't see how you can possibly know this. But, if you think you do, you might want to contact a Bayesian statistician, or read Chapter **??**."

Moyers: "Let's summarize. We now have some standards for this particular case where we take two polls and want to say whether there is been a change in support for one candidate. We should use random polls with at least 2000 voters. If the change in support level is greater than a threshold of 2.3% we are justified in reporting our results as indicating a change in support greater than 1%."

Donaldson: "But what if the measured change in support were greater than 10%. I'd feel pretty silly reporting only that the change is greater than 1%."

Statistician: "True. In fact, you could always make another null hypothesis — say, the support change is greater than 8% — and compute a p-value for your data against that null. If the p-value is low enough, you'd be justified in reporting that the change is greater than 8%. Remember, the null and alternative hypotheses here were framed for the purpose of figuring out how many people to interview in the poll. Once you have the data in hand, these hypotheses are of no particular relevance."

After a pause, the statistician adds: "Please remember that these results apply only to an increase in support for the underdog. If you want to report either an increase or a decrease, we need to do a two-tailed calculation and the sample size would need to be bigger."