3

Introduction to Hypothesis Testing: Permutation Tests

3.1 Introduction to Hypothesis Testing

Suppose scientists invent a new drug that supposedly will inhibit a mouse's ability to run through a maze. The scientists design an experiment in which three mice are randomly chosen to receive the drug and another three mice serve as controls by ingesting a placebo. The time each mouse takes to go through a maze is measured in seconds. Suppose the results of the experiment are as follows:

Drug			Control		
30	**25**	**20**	18	21	22

The average time for the drug group is 25 s, and the average time for the control group is 20.33 s. The mean difference in times is $25 - 20.33 = 4.67$ s.

The average time for the mice given the drug is greater than the average time for the control group, but this could be due to random variability rather than a real drug effect. We cannot tell for sure whether there is a real effect. What we do instead is to estimate how easily pure random chance would produce a difference this large. If that probability is small, then we conclude there is something other than pure random chance at work, and conclude that there is a real effect.

If the drug really does not influence times, then the split of the six observations into two groups was essentially random. The outcomes could just as easily be distributed.

Drug			Control		
30	**25**	18	**20**	21	22

Mathematical Statistics with Resampling and R, Second Edition. Laura M. Chihara and Tim C. Hesterberg.

In this case, the mean difference is $((30 + 25 + 18)/3) - ((20 + 21 + 22)/3) = 3.33$.

There are $\binom{6}{3} = 20$ ways to distribute 6 numbers into two sets of size 3, ignoring any ordering with each set. Of the 20 possible difference in means, 3 are as large or larger than the observed 4.67, so the probability that pure chance would give a difference this large is $3/20 = 0.15$.

Fifteen percent is small, but not small enough to be remarkable. It is plausible that chance alone is the reason the mice in the drug group ran slower (had larger times) through the maze.

For comparison, suppose a friend claims that she can control the flip of a coin, producing a head at will. You are skeptical; you give her a coin, and she indeed flips a head, three times. Are you convinced? I hope not; that could easily occur by chance, with a 12.5% probability.

This is the core idea of *statistical significance* or classical *hypothesis testing* – to calculate how often pure random chance would give an effect as large as that observed in the data, in the absence of any real effect. If that probability is small enough, we conclude that the data provide convincing evidence of a real effect.

If the probability is not small, we do not make that conclusion. This is not the same as concluding that there is no effect; it is only that the data available do not provide convincing evidence that there is an effect. In practice, there may be just too little data to provide convincing evidence. If the drug effect is small, it may be possible to distinguish the effect from random noise with 60 mice, but not 6. More flips might make your friend's claim convincing, though it would be prudent to check for a two-headed coin. (One of the authors had one, and had a former magician professor who could flip whichever side he wanted; see http://news-service.stanford.edu/news/2004/june9/diaconis-69.html.)

3.2 Hypotheses

We formalize the core idea using the language of statistical *hypothesis testing*, also known as *significance testing*.

Definition 3.1 The *null hypothesis*, denoted H_0, is a statement that corresponds to no real effect. This is the status quo, in the absence of the data providing convincing evidence to the contrary.

The *alternative hypothesis*, denoted H_A, is a statement that there is a real effect. The data may provide convincing evidence that this hypothesis is true.

A hypothesis should involve a statement about a population parameter or parameters, commonly referred to as θ; the null hypothesis is $H_0 : \theta = \theta_0$ for some θ_0. A *one-sided alternative hypothesis* is of the form $H_A : \theta > \theta_0$ or $H_A : \theta < \theta_0$; a *two-sided alternative hypothesis* is $H_A : \theta \neq \theta_0$. ||

Example 3.1 Consider the mice example in Section 3.1. Let μ_d denote the true mean time that a randomly selected mouse that received the drug takes to run through the maze; let μ_c denote the true mean time for a control mouse. Then H_0: $\mu_d = \mu_c$. That is, on average, there is no difference in the mean times between mice who receive the drug and mice in the control group.

The alternative hypothesis is H_A: $\mu_d > \mu_c$. That is, on average, mice who receive the drug have slower times (larger values) than the mice in the control group.

The hypotheses may be rewritten as H_0: $\mu_d - \mu_c = 0$ and H_A: $\mu_d - \mu_c > 0$; thus $\theta = \mu_d - \mu_c$ (any function of parameters is itself a parameter). □

The next two ingredients in hypothesis testing are a numerical measure of the effect and the probability that chance alone could produce that measured effect.

Definition 3.2 A *test statistic* is a numerical function of the data whose value determines the result of the test. The function itself is generally denoted $T = T(\boldsymbol{X})$ where $\boldsymbol{X}$ represents the data, e.g. $T = T(X_1, X_2, \ldots, X_n)$ in a one-sample problem, or $T = T(X_1, X_2, \ldots, X_m, Y_1, \ldots, Y_n)$ in a two-sample problem. After being evaluated for the sample data x, the result is called an *observed test statistic* and is written in lower-case, $t = T(x)$. ||

Definition 3.3 The *P-value* is the probability that chance alone would produce a test statistic as extreme as the observed test statistic if the null hypothesis were true. For example, if large values of the test statistic support the alternative hypothesis, the P-value is the probability $P(T \geq t)$. ||

Example 3.2 In the mice example (Section 3.1, we let the test statistic be the difference in means, $T = T(X_1, X_2, X_3, Y_1, Y_2, Y_3) = \overline{X} - \overline{Y}$ with observed value $t = \overline{x} - \overline{y} = 4.67$. Large values of the test statistic support the alternative hypothesis, so the P-value is $P(T \geq 4.67) = 3/20$. □

Definition 3.4 A result is *statistically significant* if it would rarely occur by chance. How rarely? It depends on context, but, for example, a P-value of 0.0002 would indicate that assuming the null hypothesis is true, the observed outcome would occur just 2 out of 10 000 times by chance alone, which in most circumstances seems pretty rare; you would conclude that the evidence supports the alternative hypothesis. ||

Example 3.3 Suppose public health officials are concerned about lead levels in drinking water due to old pipes throughout a city. The officials will measure lead levels in a sample of households and test the hypotheses that lead levels are at a safe level versus the alternative that the lead levels are at an unsafe level.

They collect data and find that the mean value of lead found in these households is at an unsafe level, with a P-value of 0.06. If lead levels in the city are truly safe, should we consider an outcome that occurs 6 out of 100 by chance a rare event? Considering the consequences of being wrong, officials might conclude that this result is statistically significant and something other than chance variability accounts for the mean lead level they obtained; they would conclude that lead levels in the city are indeed unsafe.

On the other hand, suppose you want to prepare for the College Board SAT Math exam. An online company provides intense tutoring at a cost of $1000. You find the results of an experiment conducted by an independent researcher that tested the hypotheses that with this tutoring, the mean SAT math score will stay the same versus the mean SAT math score will increase. From their data, they find that the mean score increases by 10 points with a P-value of 0.06. So, if the tutoring is not effective (mean score stays the same), then 6 out of 100 times, we'd obtain the observed result by chance. Is that enough evidence to convince you that the mean increase is statistically significant and it is the intense tutoring that explains the increase? At a cost of $1000, would you sign up for the tutoring? What if the cost of the tutoring was $5? □

The smaller you require the P-value to be to declare the outcome statistically significant, the more conservative you are being: You are requiring stronger evidence to reject the status quo (the null hypothesis). We will discuss P-values in more detail in Chapter 8.

Rather than just calculating the probability, we often begin by answering a larger question: What is the distribution of the test statistic when there is no real effect? For example, Table 3.1 gives all values of the test statistic in the mice example; each value has the same probability if there is no drug effect.

Definition 3.5 The *null distribution* is the distribution of the test statistic if the null hypothesis is true. ||

You can think of the null distribution as a reference distribution; we compare the observed test statistic to this reference to determine how unusual the observed test statistic is. Figure 3.1 shows the cumulative distribution function of the null distribution in the mice example.

There are different ways to calculate exact or approximate null distributions, and P-values. For now we focus on one method – permutation tests.

3.3 Permutation Tests

In the mice example in Section 3.1, we compared the test statistic to a reference distribution using permutations of the observed data. We investigate this approach in more detail.

Table 3.1 All possible distributions of {30, 25, 20, 18, 21, 22} into two sets.

Drug			Control			$\overline{X}_{\text{D}}$	$\overline{X}_{\text{C}}$	Difference in means
18	20	21	22	25	30	19.67	25.67	−6.00
18	20	22	21	25	30	20	25.33	−5.33
18	20	25	21	22	30	21	24.33	−3.33
18	20	30	21	22	25	22.67	22.67	0.00
18	21	22	20	25	30	20.33	25	−4.67
18	21	25	20	22	30	21.33	24	−2.67
18	21	30	20	22	25	23	22.33	0.67
18	22	25	20	21	30	21.67	23.67	−2.00
18	22	30	20	21	25	23.33	22	1.33
18	25	30	20	21	22	24.33	21	3.33
20	21	22	18	25	30	21	24.33	−3.33
20	21	25	18	22	30	22	23.33	−1.33
20	21	30	18	22	25	23.67	21.67	2.00
20	22	25	18	21	30	22.33	23	−0.67
20	22	30	18	21	25	24	21.33	2.67
20	25	30	18	21	22	**25**	**20.33**	**4.67** *
21	22	25	18	20	30	22.67	22.67	0.00
21	22	30	18	20	25	24.33	21	3.33
21	25	30	18	20	22	**25.33**	**20**	**5.33** *
22	25	30	18	20	21	**25.67**	**19.67**	**6.00** *

Rows where the difference in means exceeds the original value are highlighted.

Recall the beer and hot wings case study in Section 1.9. The mean number of wings consumed by females and males were 9.33 and 14.53, respectively, while the standard deviations were 3.56 and 4.50, respectively. See Figure 3.2 and Table 3.2.

The sample means for the males and females are clearly different, but the difference (14.53 − 9.33 = 5.2) could have arisen by chance. Can the difference *easily* be explained by chance alone? If not, we will conclude that there are genuine gender differences in hot wings consumption.

For a hypothesis test, let μ_{M} denote the mean number of hot wings consumed by males and μ_{F} denote the mean number of hot wings consumed by females. We test

$$H_0\colon \mu_{\text{M}} = \mu_{\text{F}} \quad \text{versus} \quad H_{\text{A}}\colon \mu_{\text{M}} > \mu_{\text{F}}$$

or equivalently

$$H_0\colon \mu_{\text{M}} - \mu_{\text{F}} = 0 \quad \text{versus} \quad H_{\text{A}}\colon \mu_{\text{M}} - \mu_{\text{F}} > 0.$$

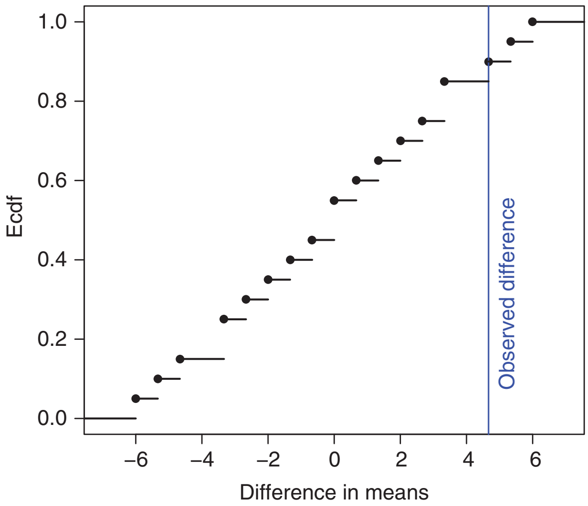

Figure 3.1 Empirical cumulative distribution function of the null distribution for difference in means for mice.

We use $T = \overline{X}_M - \overline{X}_F$ as a test statistic, with observed value $t = 5.2$.

Suppose there really is no gender influence in the number of hot wings consumed by bar patrons. Then the 30 numbers come from a single population, the way they were divided into two groups (by labeling some as male and others as female) is essentially random, and any other division is equally likely. For instance, the distribution of hot wings consumed might have been as below:

Females					Males				
5	**6**	7	7	**8**	**4**	**5**	7	8	**9**
8	**11**	**12**	**13**	**14**	11	**12**	13	13	**13**
14	14	16	16	21	17	17	18	18	21

In this case, the difference in means is 12.4 − 11.47 = 0.93.

We could proceed, as in the mice example, calculating the difference in means for *every* possible way to split the data into two samples of size 15 each. This would result in $\binom{30}{15} = 155\ 117\ 520$ differences! In practice, such exhaustive calculations are impractical unless the sample sizes are small, so we resort to sampling instead.

We create a *permutation resample*, or *resample* for short, by drawing $m = 15$ observations *without* replacement from the pooled data to be one sample (the males), leaving the remaining $n = 15$ observations to be the second sample (the females). We calculate the statistic of interest, for example, difference in means of the two samples. We repeat this many times (1000 or more). The P-value is then the fraction of times the random statistic exceeds[1] the original statistic.

[1] In hypothesis testing, "exceeds" means ≥ rather than >.

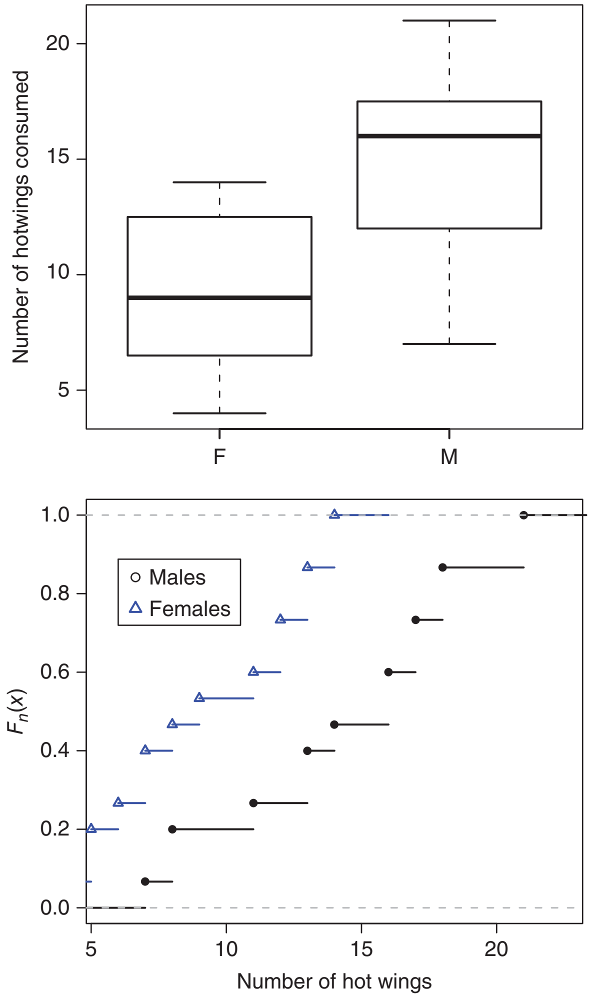

Figure 3.2 Number of hot wings consumed by gender.

Table 3.2 Hot wings consumption.

Females					Males				
4	**5**	**5**	**6**	**7**	7	8	8	11	13
7	**8**	**9**	**11**	**12**	13	14	16	16	17
12	**13**	**13**	**14**	**14**	17	18	18	21	21

We follow this algorithm:

Two-sample Permutation Test

```
Pool the $m + n$ values.
repeat
    Draw a resample of size $m$ without replacement.
    Use the remaining $n$ observations for the other sample.
    Calculate the difference in means or another statistic that compares samples.
until we have enough samples.
Calculate the $P$-value as the fraction of times the random statistics
exceed the original statistic. Multiply by 2 for a two-sided test.
Optionally, plot a histogram of the random statistic values.
```

The distribution of this difference across all permutation resamples is the *permutation distribution* (Figure 3.3). This may be exact (calculated exhaustively) or approximate (implemented by sampling). In either case, we usually use statistical software for the computations. Here is code that will perform the test in R.

R Note:

We first compute the observed mean difference in the number of hot wings consumed by males and females.

```
> tapply(Beerwings$Hotwings, Beerwings$Gender, mean)
        F        M
  9.333333 14.533333
> observed <- 14.5333 - 9.3333  # store observed mean difference
> observed
[1] 5.2
```

Since we will be working with the hot wings variable, we will create a vector holding these values. Then we will draw a random sample of size 15 from the numbers 1 through 30 (there are 30 observations total). The hot wing values corresponding to these positions will be values for the males and the remaining ones for the females. The mean difference of this permutation will be stored in `result`. This will be repeated many times.

```
hotwings <- Beerwings$Hotwings
# Another way:
# hotwings <- subset(Beerwings, select = Hotwings, drop = T)
```

```
N <- 10^5 - 1          # number of times to repeat this process
result <- numeric(N) # space to save the random differences
for (i in 1:N)
{ # sample of size 15, from 1 to 30, without replacement
  index <- sample(30, size = 15, replace = FALSE)
  result[i] <- mean(hotwings[index]) - mean(hotwings[-index])
}
```

We first create a histogram of the permutation distribution and add a vertical line at the observed mean difference.

```
hist(result, xlab = "xbar1 - xbar2",
     main = "Permutation Distribution for hot wings")
abline(v = observed, col = "blue")
# add line at observed mean difference
```

We determine how likely it is to obtain an outcome as larger or larger than the observed value.

```
> (sum(result >= observed) + 1)/(N + 1)  # P-value
[1] 0.000831                    # results will vary
```

The code snippet `result >=observed` results in a vector of `TRUE`'s and `FALSE`'s depending on whether or not the mean difference computed for a resample is greater than the observed mean difference.

`sum(result >= observed)` counts the number of `TRUE`'s. Thus, the computed P-value is just the proportion of statistics (including the original) that are as large or larger than the original mean difference.

From the output, we see that the observed difference in means is 5.2. The P-value is 0.000 831. Of the $10^5 - 1$ resamples computed by R, less than 0.1% of the resampled difference in means were as large or larger than 5.2. There are two possibilities – either there is a real difference, or there is no real effect but a miracle occurred giving a difference well beyond the range of normal chance variation. We cannot rule out the miracle, but the evidence does support the hypothesis that females in this study consume fewer hot wings than males.

The participants in this study were a convenience sample: They were chosen because they happened to be at the bar when the study was conducted. Thus, we cannot make any inference about a population.

3.3.1 Implementation Issues

We note here some implementation issues for permutation tests. The first (choice of test statistic) applies to both the exhaustive and sampling implementations, while the final three (add one to both numerator and denominator,

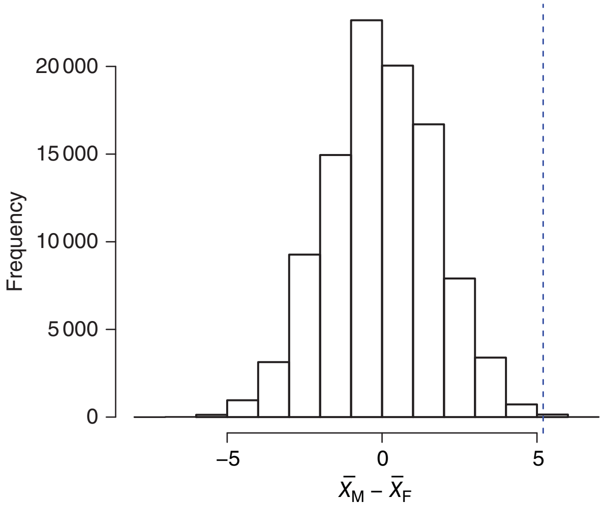

Figure 3.3 Permutation distribution of the difference in means, male–female, in the beer and hot wings example.

sample with replacement from null distribution, and more samples for better accuracy) are specific to sampling.

3.3.1.1 Choice of Test Statistic

In the examples above, we used the difference in means. We could have equally well used $\overline{X}$ (the mean of the first sample), $m\overline{X}$ (the sum of the observations in the first sample), or a variety of other test statistics. For example, in Table 3.1, the same three rows have test statistics that exceed the observed test statistic, whether the test statistic is difference in means or $\overline{X}_{\mathrm{D}}$ (the mean of the sample in the drug group).

Here is the result that states this more formally:

Theorem 3.3.1 In permutation testing, if two test statistics T_1 and T_2 are related by a strictly increasing function, $T_1(\boldsymbol{X}^*) = f(T_2(\boldsymbol{X}^*))$ where $\boldsymbol{X}^*$ is any permutation resample of the original data x, then they yield exactly the same P-values, for both the exhaustive and resampling versions of permutation testing.

Proof. For simplicity, we consider only a one-sided (greater) test. Let $\boldsymbol{X}^*$ be any permutation resample. Then

$$\begin{aligned} p_1 &= P(T_2(\boldsymbol{X}^*) \geq T_2(x)) \\ &= P(f(T_2(\boldsymbol{X})) \geq f(T_2(x))) && \text{since} f \text{ is strictly increasing} \\ &= P(T_1(\boldsymbol{X}^*) \geq T_1(x) && \text{by hypothesis.} \end{aligned}$$

Furthermore, in the sample implementation, exactly the same permutation resamples have $T_2(\boldsymbol{X}) \geq T_2(x)$ as have $T_1(\boldsymbol{X}) \geq T_1(x)$, so counting the number or fraction of samples that exceed the observed statistic yields the same results. □

Remark One subtle point is that the transformation needs to be strictly monotone *for the observed data*, not for all possible sets of data. For example, in the mice example, we used $p = P(\overline{X}_1 - \overline{X}_2 \geq \overline{x}_1 - \overline{x}_2)$. Let $T_1 = \overline{X} = \overline{X}_1 - \overline{X}_2$ denote the mean difference, and let $T_2 = \overline{X}_1$ denote the mean of just the treatment group. Let $S_1 = 3\overline{X}_1$ and $S_2 = 3\overline{X}_2$ be the sums in the two samples, and $S = S_1 + S_2 = 136$ the overall sum; this is the same for every resample (it is the sum of the same data, albeit in a different order), so we can rewrite

$$\overline{X}_2 = \frac{S_2}{3} = \frac{S - S_1}{3} = \frac{136}{3} - \overline{X}_1$$

and

$$\overline{X}_1 - \overline{X}_2 = 2\overline{X}_1 - \frac{136}{3}.$$

Hence, the transformation is $f(T_2) = 2T_2 - 136/3$. This is linear in T_2 and hence monotone (increasing). For these data, it is true that $\overline{X}_1 - \overline{X}_2 \geq 4.67$ if and only if $\overline{X}_1 \geq 25$, but that is not true for every possible set of data.

In other words, the transformation may depend on the original data; $T_1(\boldsymbol{X}^*) = f(T_2(\boldsymbol{X}^*); x)$.

3.3.1.2 Add One to Both Numerator and Denominator

When computing the P-value in the sampling implementation, we add one to both numerator and denominator. This corresponds to including the original data as an extra resample. This is a bit conservative, and avoids reporting an impossible P-value of 0.0 – since there is always at least one resample that is as extreme as the original data, namely, the original data itself.

3.3.1.3 Sample with Replacement from the Null Distribution

In the sampling implementation, we do not attempt to ensure that the resamples are unique. In effect, we draw resamples *with replacement* from the population of $\binom{m+n}{m}$ possible resamples, and hence obtain a sample with replacement from the $\binom{m+n}{m}$ test statistics that make up the exhaustive null distribution. Sampling without replacement would be more accurate, but it is not feasible, requiring too much time and memory to check that a new sample does not match any previous sample.

3.3.1.4 More Samples for Better Accuracy

In the hot wings example, we resampled 99 999 times. In general, the more resamples the better. If the true P-value is p, the estimated P-value has variance approximately equal to $p(1 - p)/N$, where N is the number of resamples.

Remark Just as the original n data values are a sample from the population, so too the N resampled statistics are a sample from a population (in this case, the null distribution). ||

The next example features highly skewed distributions and unbalanced sample sizes, as well as the need for high accuracy.

Example 3.4 Recall the Verizon case study in Section 1.3. Whether Verizon is judged to be making repairs slower for competitors' customers is determined using hypothesis tests, as mandated by the New York Public Utilities Commission (PUC). Thousands of tests are performed to compare the speed of different types of repairs, over different time periods, relative to different competitors. If substantially more than 1% of the tests give P-values below 1%, then Verizon is deemed to be discriminating.

Figure 3.4 shows the raw data for one of these tests. The mean of 1664 repairs for ILEC customers is 8.4 h, while the mean for 23 repairs for CLEC customers is 16.5 h. Could a difference that large easily be explained by chance? There

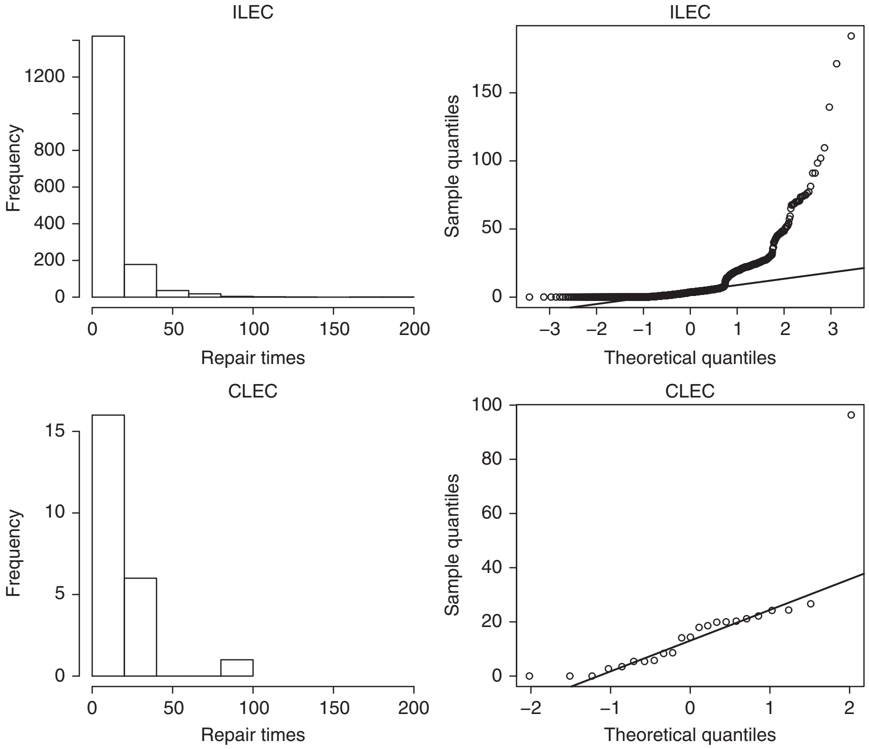

Figure 3.4 Distribution of repair times for Verizon (ILEC) and competitor (CLEC) customers. Note that the Y-axis scales are different.

appears to be one outlier in the smaller data set; perhaps that explains the difference in means? However, it would not be reasonable to throw out that observation as faulty – it is clear from the larger data set that large repair times do occur fairly frequently. Furthermore, even in the middle of both distributions, the CLEC times do appear to be longer (this is apparent in panel (b)). There are curious bends in the normal quantile plot, due to 24h cycles.

Let μ_1 denote the mean repair time for the ILEC customers and μ_2 the mean repair time for the CLEC customers. We test

$$H_0\colon \mu_1 = \mu_2 \quad \text{versus} \quad H_A\colon \mu_1 < \mu_2.$$

We use a one-sided test because the alternative of interest to the PUC is that the CLEC customers are receiving worse service (longer repair times) than the ILEC customers.

R Note

```
> tapply(Verizon$Time, Verizon$Group, mean)
 CLEC     ILEC
 16.50913 8.411611
```

We will create three vectors, one containing the times for all the customers, one with the times for just the ILEC customers, and one for just the CLEC customers.

```
Time <- Verizon$Time
#Alternatively
#Time <- subset(Verizon, select = Time, drop = TRUE)
Time.ILEC <- subset(Verizon, select = Time,
                    subset = Group == "ILEC", drop = T)
Time.CLEC <- subset(Verizon, select = Time,
                    subset = Group == "CLEC", drop = T)
```

Now we compute the mean difference in repair times and store in the vector `observed`.

```
> observed <-  mean(Time.ILEC) - mean(Time.CLEC)
> observed
[1] -8.09752
```

We will draw a random sample of size 1664 (size of ILEC group) from $1, 2, \ldots, 1687$. The times that correspond to these observations will be put in the ILEC group; the remaining times will go into the CLEC group.

```
N <- 10^4-1
result <- numeric(N)
for (i in 1:N)
{
   index <- sample(1687, size = 1664, replace = FALSE)
```

```
    result[i] <- mean(Time[index]) - mean(Time[-index])
}
```

First, plot the histogram

```
hist(result, xlab = "xbar1-xbar2",
      main = "Permutation distribution for Verizon times")
abline(v = observed, ,lty = 2, col = "blue")
```

Note that here we want to find the proportion of times the resampled mean difference is *less than or equal to* the observed mean difference.

```
(sum(result <= observed) + 1)/(N + 1)
```

One run of the simulation results in a *P*-value of 0.0165 indicating that a difference in means as small or smaller than the observed difference of −8.097 would occur less than 2% of the time if the mean times were truly equal.

In the above simulation, we used $10^4 - 1$ resamples to speed up the calculations. For higher accuracy, we should use a half-million resamples; this was negotiated between Verizon and the PUC. The goal is to have only a small chance of a test wrongly being declared significant or not, due to random sampling.

The permutation distribution is shown in Figure 3.5. The *P*-value is the fraction of the distribution that falls to the left of the observed value.

This test works fine even with unbalanced sample sizes of 1664 and 23 and even for very skewed data. The permutation distribution is skewed to

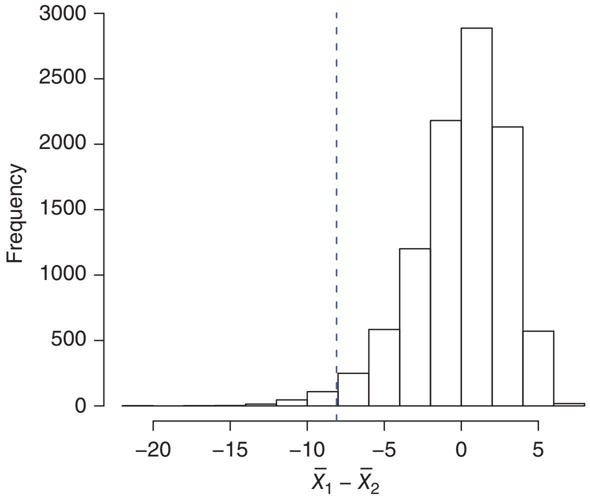

Figure 3.5 Permutation distribution of difference of means (ILEC−CLEC) for the Verizon repair time data.

the left, but that doesn't matter; both the observed statistic and the permutation resamples are affected by the size imbalance and skewness in the same way. □

3.3.2 One-sided and Two-sided Tests

For the hypothesis test with alternative $H_A : \mu_1 - \mu_2 < 0$, we compute a P-value by finding the fraction of resample statistics that are less than or equal to the observed test statistic (or greater than or equal to for the alternative $\mu_1 - \mu_2 > 0$.)

For a two-sided test, we calculate both one-sided P-values, multiply the smaller by 2, and finally (if necessary) round down to 1.0 (because probabilities can never be larger than 1.0).

In the mice example with observed test statistic $t = 4.67$, the one-sided P-values are 3/20 for H_A: $\mu_d - \mu_c > 0$ and 18/20 for H_A: $\mu_d - \mu_c < 0$. Hence the two-sided P-value is $6/20 = 0.30$ (recall Table 3.1).

Two-sided P-values are the default in statistical practice – you should perform a two-sided test unless there is a clear reason to pick a one-sided alternative hypothesis. It is not fair to look at the data before deciding to use a one-sided hypothesis.

Example 3.5 We return to the `Beerwings` data set, and the comparison of the mean number of hot wings consumed by males and females. Suppose prior to this study, we had no preconceived idea of which gender would consume more hot wings. Then our hypotheses would be

$$H_0 : \mu_M = \mu_F \quad \text{versus} \quad H_A : \mu_M \neq \mu_F.$$

We found the one-sided P-value (for alternative "greater") to be 0.000 831, so for a two-sided test, we double 0.000 831 to obtain the P-value 0.00 166.

If gender does not influence average hot wings consumption, a difference as extreme or more extreme than what we observed would occur only about 0.2% of the time. We conclude that males and females do not consume, on average, the same number of hot wings. □

3.3.2.1 To Obtain *P*-values in the Two-sided Case We multiply by 2

We multiply the smaller of the one-sided P-values by 2, using the observed test statistic. Multiplying by 2 has a deeper meaning. Because we are open to more than one alternative to the null hypothesis, it takes stronger evidence for any one of these particular alternatives to provide convincing evidence that the null hypothesis is incorrect. With two possibilities, the evidence must be stronger by a factor of 2, measured on the probability scale.

3.3.3 Other Statistics

We noted in Section 3.3.1 the possibility of using a variety of statistics and getting equivalent results, provided the statistics are related by a monotone transformation.

Permutation testing actually offers considerably more freedom than that; the basic procedure works with any test statistic. We compute the observed test statistic, resample, compute the test statistics for each resample, and compute the P-value (see the algorithm in Section 3.3.) Nothing in the process requires that the statistic be a mean or equivalent to a mean.

This provides the flexibility to choose a test statistic that is more suitable to the problem at hand. Rather than using means, for example, we might base the test statistic on *robust statistics,* that is, statistics that are not sensitive to outliers. Two examples of robust statistics are the median and the trimmed mean. We have already encountered the median. The trimmed mean is just a variant of the mean: we sort the data, omit a certain fraction of the low and high values, and calculate the mean of the remaining values. In addition, permutation tests could also compare proportions or variances. We give examples of each of these cases next, then turn in the next section to what appears at first glance to be a completely different setup but is in fact just another application of this idea.

Example 3.6 In the Verizon example we observed that the data have a long tail – there are some very large repair times (Figure 3.4). We may wish to use a test statistic that is less sensitive to these observations. There are a number of reasons we might do this. One is to get a better measure of what is important in practice and how inconvenienced customers are by the repairs. After a while, each additional hour probably does not matter as much, yet a sample mean treats an extra 10h on a repair time of 100h the same as an extra 10h on a repair time of 1h. Second, a large recorded repair time might just be a blunder; for example, a repair time of 10^6 h must be a mistake. Third, a more robust statistic could be more sensitive at detecting real differences in the distributions – the mean is so sensitive to large observations that it pays less attention to moderate observations, whereas a statistic more sensitive to moderate observations could detect differences between populations that show up in the moderate observations.

Here is the R code for permutation tests using medians and trimmed means (Figure 3.6)

R Note for Verizon, cont.

```
observed <- median(Time.ILEC) - median(Time.CLEC)
N <- 10^4-1
```

```
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(1687, size = 1664, replace = FALSE)
  result[i] <- median(Time[index]) - median(Time[-index])
}
(sum(result <= observed) + 1)/(N + 1)  # P-value
```

To obtain the results for the trimmed mean, we add the option `trim=.25` to the `mean` command. Substitute the following in the above:

```
observed  <- (mean(Time.ILEC, trim = .25) -
               mean(Time.CLEC, trim = .25))
result[i] <- (mean(Time[index], trim = .25) -
               mean(Time[-index], trim = .25))
```

It seems apparent that these more robust statistics are more sensitive to a possible difference between the two populations; the tests are significant with estimated *P*-values of 0.002 and 0.001, respectively. The figures (Figure 3.6) also suggest that the observed statistics are well outside the range of normal chance variation.

One caveat is in order – it is wrong to try many different tests, possibly with minor variations, until you obtain a statistically significant outcome. If you try enough different things, eventually one will come out significant, whether or not there is a real difference.

There are ways to guard against this and in Section 8.5.3, we will learn about different corrections to avoid these false positives.

We can also apply permutation tests to questions other than comparing the centers of two populations, for example the difference between the two populations in the proportion of repair times that exceed 10h or the ratio of variances of the two populations. Using the R code below, it appears that the proportions do differ (*P*-value = 0.0008, one sided), while the variances do not (*P*-value = 0.258, two sided). The permutation distributions are very different (see Figure 3.7), but this does not affect the validity of the method.

R Note for Verizon, cont.

We will first create two vectors that will contain the repair times for the ILEC and CLEC customers, respectively. The command `mean(Time.ILEC > 10)` computes the proportion of times the ILEC times are greater than 10.

```
> observed <- mean(Time.ILEC > 10) - mean(Time.CLEC > 10)
> observed
[1] -0.336852
```

Thus, about 33.7% fewer ILEC customers had repair times exceeding 10 h.
We reuse the previous code for trimmed means but with the following modification that computes the difference in proportions:

```
result[i] <- mean(Time[index]>10) - mean(Time[-index] > 10)
```

To perform the test for the ratio of variances substitute:

```
observed  <- var(Time.ILEC)/var(Time.CLEC)
result[i] <- var(Time[index])/var(Time[-index])
```

□

3.3.4 Assumptions

Under what conditions can we use the permutation test? First, the permutation test makes no distributional assumption on the two populations under consideration. That is, there is no requirement that samples are drawn from a normal distribution, for example.

In fact, permutation testing does not even require that the data be drawn by random sampling from two populations. A study for the treatment of a rare disease could include all patients with the disease in the world. In this case, it does require that subjects be assigned to the two groups randomly.

In the usual case that the two groups are samples from two populations, pooling the data does require that the two *populations* have the same distribution when the null hypothesis is true. They must have the same mean, spread, and shape. This does not mean that the two *samples* must have the same mean, spread, and shape – there will always be some chance variation in the data.

In practice, the permutation test is usually robust when the two populations have different distributions. The major exception is when the two populations have different spreads, and the sample sizes are dissimilar. This exception is rarely a concern in practice, unless you have other information (besides the data) that the spreads are different. For example, one of us consulted for a large pharmaceutical company testing a new procedure for measuring a certain quantity; the new procedure was substantially cheaper, but not as accurate. The loss of accuracy was acceptable, provided that the mean measurements matched. This is a case where permutation testing would be doubtful, because it would pool data from different distributions. Even then, it would usually work fine if the sample sizes were equal.

Example 3.7 We investigate the extreme case in more detail. Suppose population A is normal with mean 0 and variance $\sigma_{\mathrm{A}}^2 = 10^6$, and population B is normal with mean 0 and variance $\sigma_{\mathrm{B}}^2 = 1$. Draw a sample of size $n_{\mathrm{A}} = 10^2$ from population A and a sample of size $n_{\mathrm{B}} = 10^6$ from population B. Thus, we have

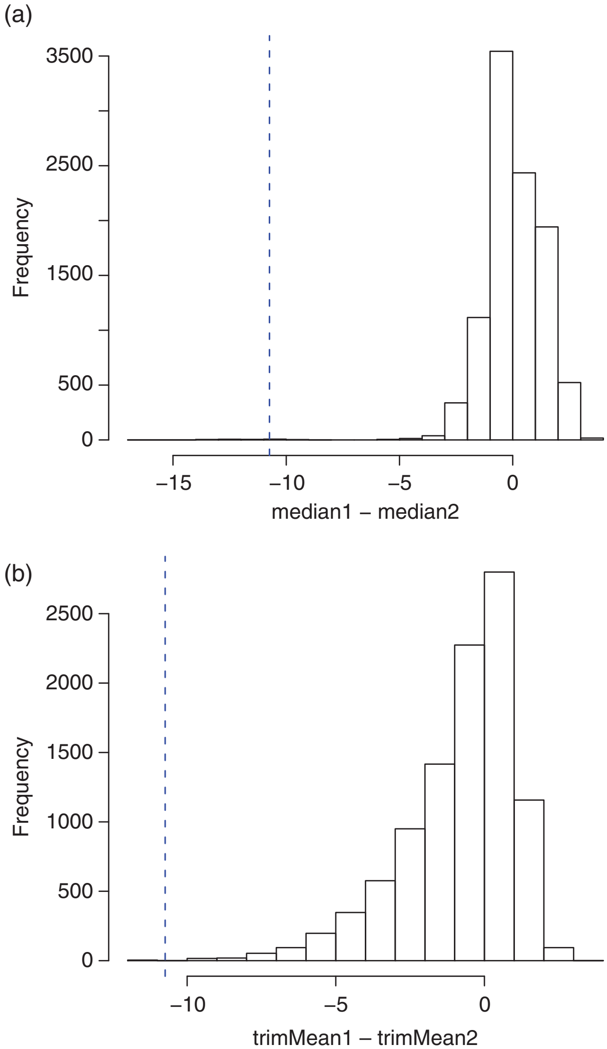

Figure 3.6 Repair times for Verizon data. (a) Permutation distribution for difference in medians. (b) Permutation distribution for difference in 25% trimmed means.

that the null hypothesis is true with both populations having mean 0. Let the test statistic be $T = \overline{X}_{\mathrm{A}}$. When drawing the original sample, T has variance $\sigma_{\mathrm{A}}^2/n_{\mathrm{A}} = 10^4$ (by Theorem A.4.1). What is the probability that this statistic T is greater than, say, 5? By standardizing, we find

$$P(T \geq 5) = P\left(\frac{T}{100} \geq \frac{5}{100}\right) = P(Z \geq 0.05) = 0.48.$$

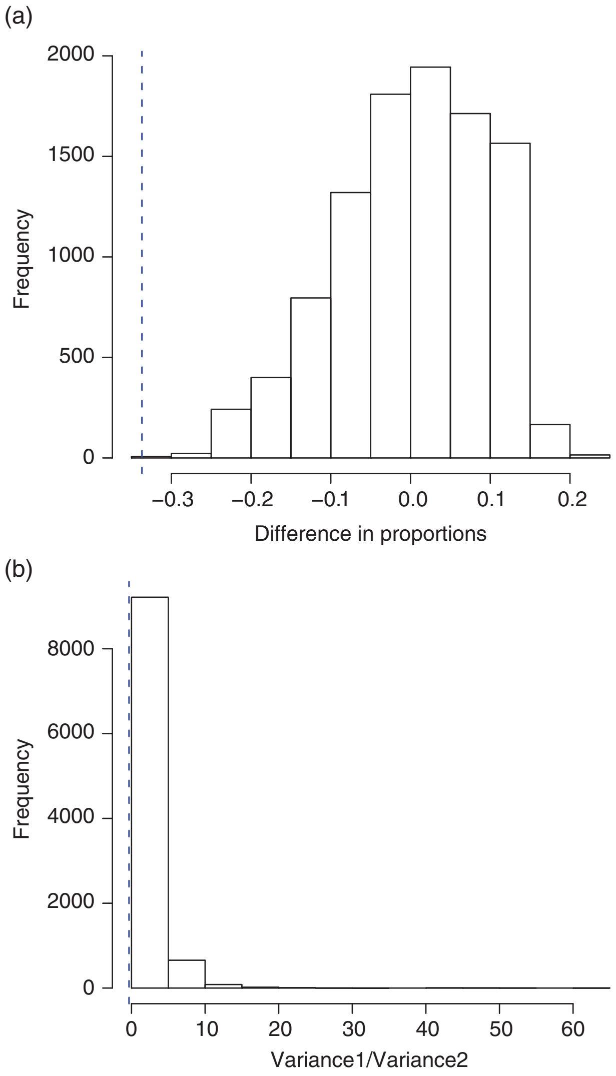

Figure 3.7 Repair times for Verizon data. (a) Difference in proportion of repairs exceeding 10 h. (b) Ratio of variances (ILEC/CLEC).

Thus, with its huge variance of 10^4, there is nearly a 50% chance of T being greater than 5.

When we pool the two samples, it turns out that the variance of the permutation distribution of T is around $(n_\mathrm{A}\sigma_\mathrm{A}^2 + n_\mathrm{B}\sigma_B^2)/(n_\mathrm{A} + n_\mathrm{B}) \approx 101$ (plus or minus random variation). Thus, when we perform the permutation test, the resampled

T's have variance around $101/n_{\mathrm{A}} \approx 1.01$, or equivalently, a standard deviation about 1.005 (again, by Theorem A.4.1). So almost none of the permutation T's will be larger than 5:

$$P(T \geq 5) = P\left(\frac{T}{1.005} \geq \frac{5}{1.005}\right) = P(Z \geq 4.975) = 0.$$

Thus, there is nearly a 50% chance of reporting a P-value near 0 and erroneously concluding that the means are not the same. □

Example 3.8 In the Iowa recidivism case study in Section 1.4, we have the population of offenders, convicted in Iowa of either a felony or misdemeanor, who were released from prison in 2010. Of these, 36.5% of those under 25 years of age were sent back to prison compared with 30.6% of those 25 years of age or older, so the observed difference in proportions is 0.059. Is this a statistically significant difference? We can perform a permutation test to check.

R Note

The variable `Recid` in `Recidivism` is a factor variable with two levels, "Yes" and "No." We use `ifelse` to convert this to a numeric binary variable.

```
k <- complete.cases(Recidivism$Age25)    #omit NA's
Recid2 <- ifelse(Recidivism$Recid[k] == "Yes", 1, 0)
```

There were 3077 offenders under the age of 25 and 13942 offenders who were 25 years of age or older.

```
observed <- .365 - .306
for (i in 1:N)
{
index <- sample(17019, size = 3077, replace = FALSE)
result[i] <- mean(Recid2[index]) - mean(Recid2[-index])
}

2*(sum(result >= observed)+1)/(N+1)
```

For a two-sided test, the P-value is 2×10^{-5}, so we conclude that there is a statistically significant difference in recidivism between those under 25 years of age and those 25 years of age or older.

As we noted, the permutation test is applicable when the data are a population as opposed to a sample from a population. This test tells us that if recidivism was a random occurrence, unrelated to age group, then the chance of observing an outcome as extreme or more extreme than the observed difference in proportions of 0.059 is 2×10^{-5}. □

Table 3.3 Partial view of `Beerwings` data set.

	Gender	Hot wings			Gender	Hot wings
1	F	4		11	F	9
2	F	5		26	F	17
3	F	5		25	F	17
4	F	6		2	F	5
5	F	7		4	F	6
6	F	7	⇒	8	F	8
7	M	7		3	M	5
8	F	8		20	F	14
9	M	8		10	M	8
10	M	8		18	M	13
	⋮				⋮	

The `Gender` column is held fixed and the rows of the `Hotwings` variable are permuted. The first column indicates which rows of the hot wing values were permuted.

3.3.5 Remark on Terminology

Why is the two-sample permutation test above called *permutation* testing? It seems like all we are doing is splitting the data into two samples, with no hint of a permutation. Well, imagine storing the data in a table with two columns and $m + n$ rows; the first column contains labels, for example, m copies of "M" and n copies of "F," while the second contains the numerical data. We may permute the rows of either column, randomly; this is equivalent to splitting the data into two groups randomly.

Table 3.3 illustrates one such permutation of one of the columns in the `beerwings` data.

That idea of permuting the rows of one column generalizes to other situations, including the analysis of contingency tables, which we will encounter in Chapter 10.

3.4 Matched Pairs

Divers competing in the FINA 2017 World Championships perform five dives in each of several rounds.[2] The sum of the scores of these five dives determines who moves on to the next round. Do divers tend to get the same score, on average, in the semifinal and final rounds of a competition? Or might the scores in

[2] Fédération Internationale de Natation.

Table 3.4 Partial view of diving scores in file `Diving2017`.

Name	Country	Semifinal	Final
Cheog Jun Hoong	Malaysia	325.5	397.5
Si Yajie	China	382.8	396.00
Ren Qian	China	367.5	391.95
Kim Mi Rae	North Korea	346.00	385.55
⋮			

the final round be different, due to fatigue, or heightened effort, or a strategy to perform more difficult dives in the final round? We have the scores from the semifinal and final round of the 10 m platform for the top 12 female divers (Table 3.4). The average score in the semifinal is 338.50 and the final is 350.475. Is this a real difference or could this be attributed to chance variability?

Now, it may be tempting to proceed as we did in investigating the mean number of hot wings consumed by men and women, by comparing the mean scores in the semi-final and final rounds. But note that the data here are *not independent*! The scores that any particular diver receives in the semifinal and final rounds are related, in the sense that how well she dives depends on her training and her genetics. Thus, the data are called *matched pairs* or *paired data*.

So, for instance, if there is no true difference in how Qian Ren of China performs in the last two rounds, then the fact that she received a score of 367.5 in the semi-final and a 391.95 in the final is due to chance. In another circumstance, she might have received the 391.95 in the semifinal and the 367.5 in the final. For a permutation test, we randomly select some of the divers and transpose their two scores, leaving the other divers scores the same.

R Note

Since the effect of transposing the semi-final and final score for a diver results in a sign change in the difference, we will draw 12 random values from $\{-1, 1\}$. A draw of -1 indicates to transpose and multiply the difference by -1, while a 1 keeps the original order and value.

```
Diff <- Diving2017$Final - Diving2017$Semifinal  #difference in two scores
observed <- mean(Diff)                           #mean of difference

N <- 10^5-1
result <- numeric(N)

for (i in 1:N)
```

```
{
   Sign <- sample(c(-1,1), 12, replace=TRUE) #random vector of 1's or -1's
   Diff2 <-  Sign*Diff                       #random pairs (a-b) -> (b-a)
   result[i] <- mean(Diff2)                  #mean of difference
}

hist(result)
abline(v=mean(observed), col = "blue")

2*(sum(result >= observed + 1)/(N+1))        #P-value
```

We obtain a P-value of 0.21, which suggests that chance alone might account for the difference we observed in the mean diving scores in the semifinal and final rounds.

If we had performed a permutation test assuming that the final scores were independent of the semifinal scores, we would have obtained (in one simulation) a P-value of 0.165, a slightly smaller probability. Although in this example we would have reached the same conclusion, it is possible in other settings that the two approaches might lead to two conflicting outcomes. Thus, when you have two variables, it is important to think carefully about whether or not these represent data from two independent populations.

Exercises

3.1 Suppose you conduct an experiment and inject a drug into three mice. Their times for running a maze are 8, 10, and 15 s; the times for two control mice are 5 and 9 s.

a) Compute the difference in mean times between the treatment group and the control group.

b) Write out all possible permutations of these times to the two groups and calculate the difference in means.

c) What proportion of the differences are as large or larger than the observed difference in mean times?

d) For each permutation, calculate the mean of the treatment group only. What proportion of these means are as large or larger than the observed mean of the treatment group?

3.2 Your statistics professor comes to class with a big urn that she claims contains 9999 blue marbles and 1 red marble. You draw out one marble at random and finds that it is red. Would you be willing to tell your professor that you think she is wrong about the distribution of colors? Why or why not? What are you assuming in making your decision? What if instead, she claims there are nine blue marbles and 1 red one (and you draw out a red marble)?

3.3 In a hypothesis test comparing two population means, $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 > \mu_2$:

a) Which P-value, 0.03 or 0.006 provides stronger evidence for the alternative hypothesis?

b) Which P-value, 0.095 or 0.04 provides stronger evidence that chance alone might account for the observed result?

3.4 In the algorithms for conducting a permutation test, why do we add 1 to the number of replications N when calculating the P-value?

3.5 In the flight delays case study in Section 1.1, the data contain flight delays for two airlines, American Airlines and United Airlines.

a) Conduct a two-sided permutation test to see if the difference in mean delay times between the two carriers are statistically significant.

b) The flights took place in May and June of 2009. Conduct a two-sided permutation test to see if the difference in mean delay times between the two months is statistically significant.

3.6 In the flight delays case study in Section 1.1, the data contains flight delays for two airlines, American and United.

a) Compute the proportion of times that each carrier's flights was delayed more than 20 min. Conduct a two-sided test to see if the difference in these proportions is statistically significant (see the `R` Note in Example 3.6).

b) Compute the variance in the flight delay lengths for each carrier. Conduct a test to see if the variance for United Airlines differs from that of American Airlines.

3.7 In the flight delays case study in Section 1.1, repeat Exercise 3.5 part (a) using three test statistics, (i) the mean of the United Airlines delay times, (ii) the sum of the United Airlines delay times, and (iii) the difference in means, and compare the P-values. Make sure all three test statistics are computed within the same `for` loop. What do you observe?

3.8 In the flight delays case study in Section 1.1,

a) Find the trimmed mean of the delay times for United Airlines and American Airlines.

b) Conduct a two-sided test to see if the difference in trimmed means is statistically significant.

3.9 In the flight delays case study in Section 1.1,

a) Compute the proportion of times the flights in May and in June were delayed more than 20 min, and conduct a two-sided test to see if the difference between months is statistically significant.

b) Compute the ratio of the variances in the flight delay times in May and in June. Is this evidence that the true ratio is not equal to 1, or could this be due to chance variability? Conduct a two-sided test to check.

3.10 In the black spruce case study in Section 1.10, seedlings were planted in plots that were either subject to competition (from other plants) or not. Use the data set `Spruce` to conduct a test to see if the mean difference in how much the seedlings grew (in height) over the course of the study under these two treatments is statistically significant.

3.11 The file `Phillies2009` contains data from the 2009 season for the baseball team the Philadelphia Phillies.

a) Compare the empirical distribution functions of the number of strike-outs per game (`StrikeOuts`) for games played at home and games played away (`Location`).

b) Find the mean number of strike-outs per game for the home and the away games.

c) Perform a permutation test to see if the difference in means is statistically significant.

3.12 In the Iowa recidivism case study in Section 1.4, offenders had originally been convicted of either a felony or misdemeanor.

a) Use `R` to create a table displaying the proportion of felons who recidivated and the proportion of those convicted of a misdemeanor who recidivated.

b) Determine whether or not the difference in recidivism proportions computed in (a) is statistically significant.

3.13 In the Iowa recidivism case study in Section 1.4, for those offenders who recidivated, we have data on the number of days until they reoffended. For those offenders who did recidivate, determine if the difference in the mean number of days (`Days`) until recidivism between those under 25 years of age and those 25 years of age and older is statistically significant.

Remark: Data on recidivism were collected for only 3 years from time of release from prison since studies suggest that most relapses occur within that time period. Thus, it is possible that some offenders who had not relapsed in that time period, might be convicted of another crime at a later point in time. The variable `Days` is *right censored*.

3.14 Does chocolate ice cream have more calories than vanilla ice cream? The data set `IceCream` contains calorie information for a sample of brands of chocolate and vanilla ice cream.

a) Inspect the data set, then explain why this is an example of matched pairs data.

b) Compute summary statistics of the number of calories for the two flavors.

c) Conduct a permutation test to determine whether or not chocolate ice cream has, on average, more calories than vanilla ice cream.

3.15 Is there a difference in the price of groceries sold by the two retailers Target and Walmart? The data set `Groceries` contain a sample of grocery items and their prices advertised on their respective web sites on one specific day.

a) Inspect the data set, then explain why this is an example of matched pairs data.

b) Compute summary statistics of the prices for each store.

c) Conduct a permutation test to determine whether or not there is a difference in the mean prices.

d) Create a histogram of the difference in prices. What is unusual about Quaker Oats Life cereal?

e) Redo the hypothesis test without this observation. Do you reach the same conclusion?

3.16 In the sampling version of permutation testing, the one-sided P-value is $\hat{P} = (X + 1)/(N + 1)$, where X is the number of permutation test statistics that are as large or larger than the observed test statistic. Suppose the true P-value (for the exhaustive test, conditional on the observed data) is p.

a) What is the variance of $\hat{P}$?

b) What is the variance of $\hat{P}_2$ for the two-sided test (assuming that p is not close to 0.5, where p is the smaller true one-sided P-value?)