consistent with sampling variability. When constructing confidence intervals, the `resampling()` function was used. Re-sampling will typically repeat some cases and omit others. Here, the `shuffle()` function will be used instead, to scramble the order of one or more variables while leaving the others in their original state.

DATA FILE
galton.csv

To illustrate, consider a model for exploring whether `sex` and `mother`'s height are related to the height of the child:

```
> galton = fetchData("galton.csv")
> mod = lm(height ~ sex + mother, data=galton)
> coefficients(mod)
(Intercept)        sexM      mother
     41.450       5.177       0.353
```

The coefficients indicate that typical males are taller than typical females by about 5 inches and that for each inch taller the mother is, a child will typically be taller by 0.35 inches. A reasonable test statistic to summarize the whole model is R^2. The summary report shows that to be $R^2 = 0.5618$.

For confidence intervals, re-sampling was applied to the entire data frame. This selects random cases, but each selected case is an authentic one that matches exactly the original values for that case. The point of re-sampling is to get an idea of the variability introduced by random sampling of authentic cases.

```
> do(5) * lm(height ~ sex + mother, data=resample(galton))
  Intercept sexM mother sigma r-squared
1      43.6 5.16  0.319  2.27     0.578
2      41.6 5.18  0.353  2.28     0.583
3      42.8 5.00  0.333  2.33     0.552
4      35.0 5.32  0.449  2.43     0.583
5      42.2 5.15  0.341  2.39     0.552
```

The `sex`**M** coefficients are tightly grouped near 5 inches, the `mother` coefficients are around 0.3 to 0.4.

In order to carry out a permutation test, do not randomize the whole data frame. Instead, shuffle just the response variable:

```
> do(5) * lm(shuffle(height) ~ sex + mother, data=galton)
  Intercept   sexM  mother sigma r-squared
1      62.2  0.434  0.0681  3.58  0.005407
2      65.9 -0.393  0.0162  3.58  0.003151
3      62.1  0.546  0.0686  3.57  0.007513
4      69.1  0.050 -0.0373  3.59  0.000637
5      71.2 -0.448 -0.0661  3.58  0.005537
```

Now the `sex`**M** and `mother` coefficients are close to zero, as would be expected when there is no relationship between the response variable and the explanatory variables.

In constructing the sampling distribution under the null hypothesis, you should do hundreds of trials of fitting the model to the scrambled data, calculating the

test statistic (R^2 here) for each trial. Note that each trial itself involves all of the cases in your sample, but those cases have been changed so that the shuffled variable almost certainly takes on a different value in every case than in the original data.

```
> nulltrials = do(500) * lm(shuffle(height) ~ sex + mother, data=galton)
```

Notice that `do()` calculates R^2 from the model. The output of `do()` is a data frame:

```
> nulltrials

  Intercept   sexM    mother sigma r-squared
1      67.9 -0.385 -0.014799  3.58  0.002943
2      66.7  0.101  0.000573  3.59  0.000198
3      67.9  0.302 -0.020769  3.58  0.002000
... for 500 cases altogether ...
```

Naturally, all of the R^2 values for the trials are close to zero. After all, there is no relation between the response variable (after randomization with `shuffle()`) and the explanatory variables.

The p-value can be calculated directly from the trials, by comparison to the observed value in the actual data: R^2 was 0.5618.

```
> tally(~ r.squared > 0.5618, nulltrials)

 TRUE FALSE Total
    0   500   500
```

None of the 500 trials were greater than the value of the test statistic, 0.5618. It wouldn't be fair to claim that $p = 0$, since we only did 500 trials, but it is reasonable to say that the permutation test shows the p-value is $p \leq 1/500$.

14.5.2 First-Principle Tests

On modern computers, the permutation test is entirely practical. But a few decades ago, it was not. Great creativity was applied to finding test statistics where the sampling distribution could be estimated without extensive calculation. One of these is the F statistic. This is still very useful today and is a standard part of the regression report in many statistical packages.

Here is the regression report from the `height` ~ `sex`+`mother`:

```
> mod = lm( height ~ sex + mother, data=galton)
> summary(mod)

...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.4495     2.2095    18.8   <2e-16 ***
sexM          5.1767     0.1587    32.6   <2e-16 ***
mother        0.3531     0.0344    10.3   <2e-16 ***
```