Permutation procedure

To determine the possible values of D that may arise when there is no association we will use the **permutation procedure**.

- Create an artificial **permutation** dataset where the observed values of x and y are *randomly* paired together. By design, there is no association in this dataset.
- Calculate D for this permutation sample and record it.
- Repeat this process a "lot" of times and make a histogram of the values of D to see what happens "by chance" when the variables have no association.

The distribution of possible values of D that arise from this simulation (i.e., through natural variation when no association exists) is called the **sampling distribution** of D.

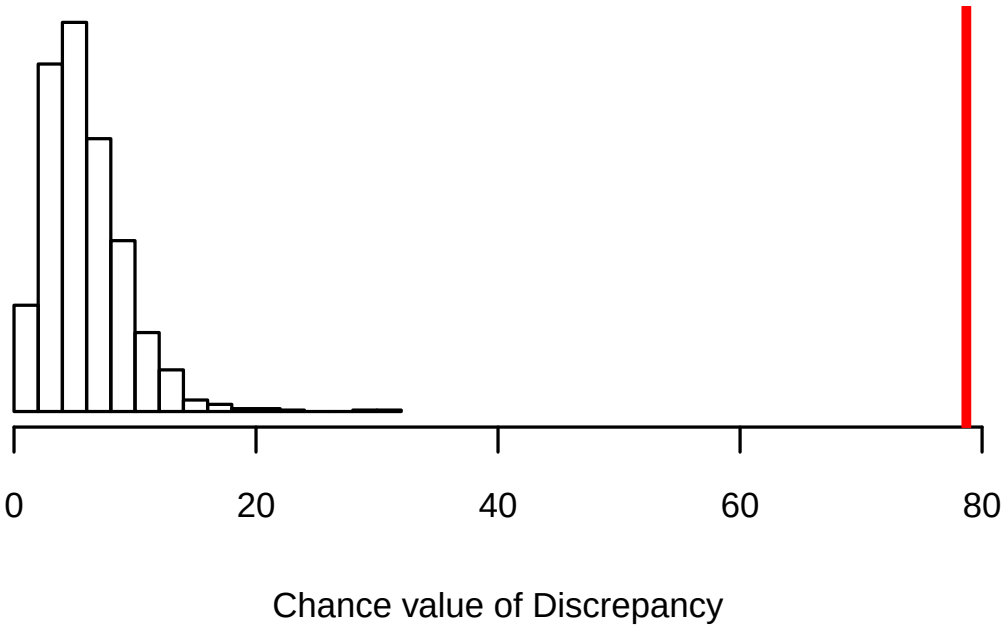
Permutation sample

The **permutation procedure** is most easily illustrated when x and y are both numerical. Below is a table of the original observed values and of three permutation datasets where the values of x and y have been randomly paired together (values in each sample have been sorted on x).

Individual	Observed Data		Permutation 1		Permutation 2		Permutation 3	
	x	y	x	y	x	y	x	y
1	2.2	0.2	2.2	3.4	2.2	5.6	2.2	3.4
2	2.7	3.4	2.7	6.0	2.7	3.4	2.7	0.2
3	2.8	3.4	2.8	5.6	2.8	6.0	2.8	5.6
4	4.9	5.6	4.9	3.4	4.9	3.4	4.9	3.4
5	7.2	6.0	7.2	0.2	7.2	0.2	7.2	6.0

Simulation results for dropped call data

The histogram shows the values of D observed over 1000 permutation datasets. The value of D observed in the data is 78.7, which is *highly* out of line with what occurs "by chance".



The discrepancy, and thus the association, is "statistically significant" since it is unlikely to have arisen by chance.

p-value of discrepancy

To perform the statistical analysis, we calculate the p-value of the association.

> The **p-value** of the observed discrepancy D is the probability that a permutation sample (created by randomly pairing x and y together so that they do not have an association by design) would produce a value of D *at least as large as what was observed in the original data*.

In other words, the p-value of the association is the probability of seeing at least as large a discrepancy between the conditional distributions of y and the marginal distribution of y "by chance" alone.

Statistical Significance

If the p-value is less than 5%, we say that the association is **statistically significant**

- Such a large discrepancy is unlikely to occur by chance (i.e., it happens less than 5% of the time).
- Note: a small p-value does not mean the association is strong.

If the p-value is 5% or greater, we say that the association is **not statistically significant**.

- An observed discrepancy of this magnitude happens "all the time" (with a greater than 5% chance) when variables do not have an association.
- There is insufficient evidence to suggest the variables are related.
- Note: a large p-value does not mean an association for sure does not exist; the association may have been too weak to be detected by the data.

Calculation of p-value

The p-value of the discrepancy D can be approximated by the permutation procedure

- Count up the number of permutation datasets which had a discrepancy of D or greater.
- Divide this by the number of permutations datasets that were generated.

For example, imagine you observed a value of $D = 51.1$ from your data and you made 50,000 permutation datasets. If you find that 23 of them had a value of $D \geq 51.1$, the approximate p-value would be $23/50000 = 0.00046$.

The test in R

We will use the (custom) command `associate()` to perform the test. You will have to load up library `regclass` first.

```
associate(y~x,data=...,permutations=500,seed=...)
```

- `y` and `x` are the column names in the data frame
- fill in `data=` with the name of the data frame. This argument can be omitted if you defined x and y manually using the left arrow convention.
- `permutations` gives the number of permutation datasets to produce. If the argument is omitted, 500 will be made.
- `seed` is an optional argument that provides the random number seed. Since the p-value is approximated by randomly pairing x and y values, it can/will differ if you run the command again. Setting `seed` to any positive integer will allow you to reproduce the results.

Example: dropped call data

Let us make 1000 permutation datasets and, for reproducibility, set the random number seed (you will get the exact same results if you)

```
associate(DropCallFreq~Provider,data=CALLS,permutations=1000,seed=2015)
```

Example: dropped call data output 1

The text output sent to the Console gives you the contingency table (observed counts) and the table of expected counts if x and y did not have an association.

```
Association between Provider (categorical) and  DropCallFreq (categorical) usin
Contingency table:
          y
x          Occasionally Often Rarely Total
  ATT                57    23     93   173
  Sprint              9     2     30    41
  USCellar           11     6     38    55
  Verizon            28     6    276   310
  Total             105    37    437   579

   Table of Expected Counts:
         Occasionally Often Rarely
ATT              31.4  11.1  130.6
Sprint            7.4   2.6   30.9
USCellar         10.0   3.5   41.5
Verizon          56.2  19.8  234.0
```

Example: dropped call data output 2

The text output sent to the Console gives you the conditional/marginal distributions.

```
Conditional distributions of y (DropCallFreq) for each level of x (Provider):
If there is no association, these should look similar to each other and
 similar to the marginal distribution of y
           Occasionally      Often    Rarely
ATT          0.32947977 0.13294798 0.5375723
Sprint       0.21951220 0.04878049 0.7317073
USCellular   0.20000000 0.10909091 0.6909091
Verizon      0.09032258 0.01935484 0.8903226
Marginal     0.18134715 0.06390328 0.7547496
```

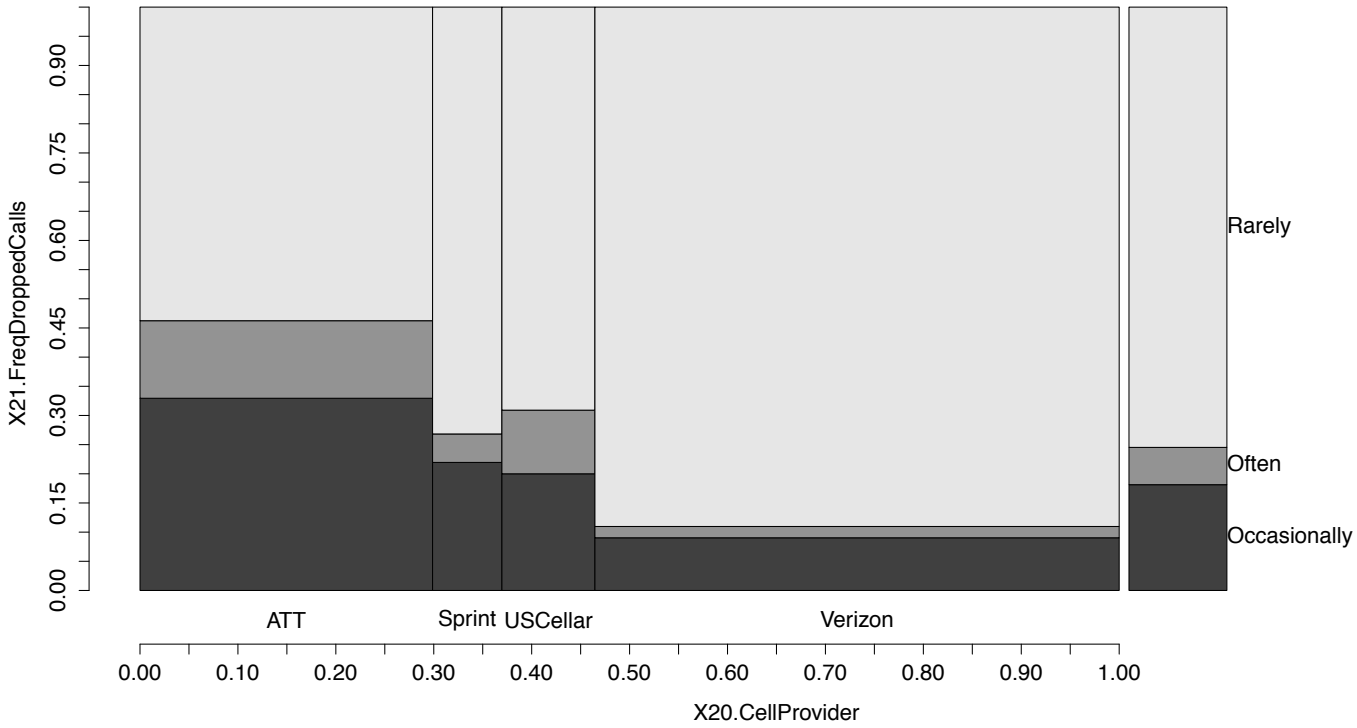
Example: dropped call data output 3

The text output sent to the Console gives you the discrepancy between observed and expected counts and the approximate p-value as found with the permutation procedure.

```
Permutation procedure:
  Discrepancy Estimated p-value
     78.65499                 0
With 1000 permutations, we are 95% confident that:
 the p-value is between 0 and 0.004
If 0.05 is in this range, change permutations= to a larger number
```

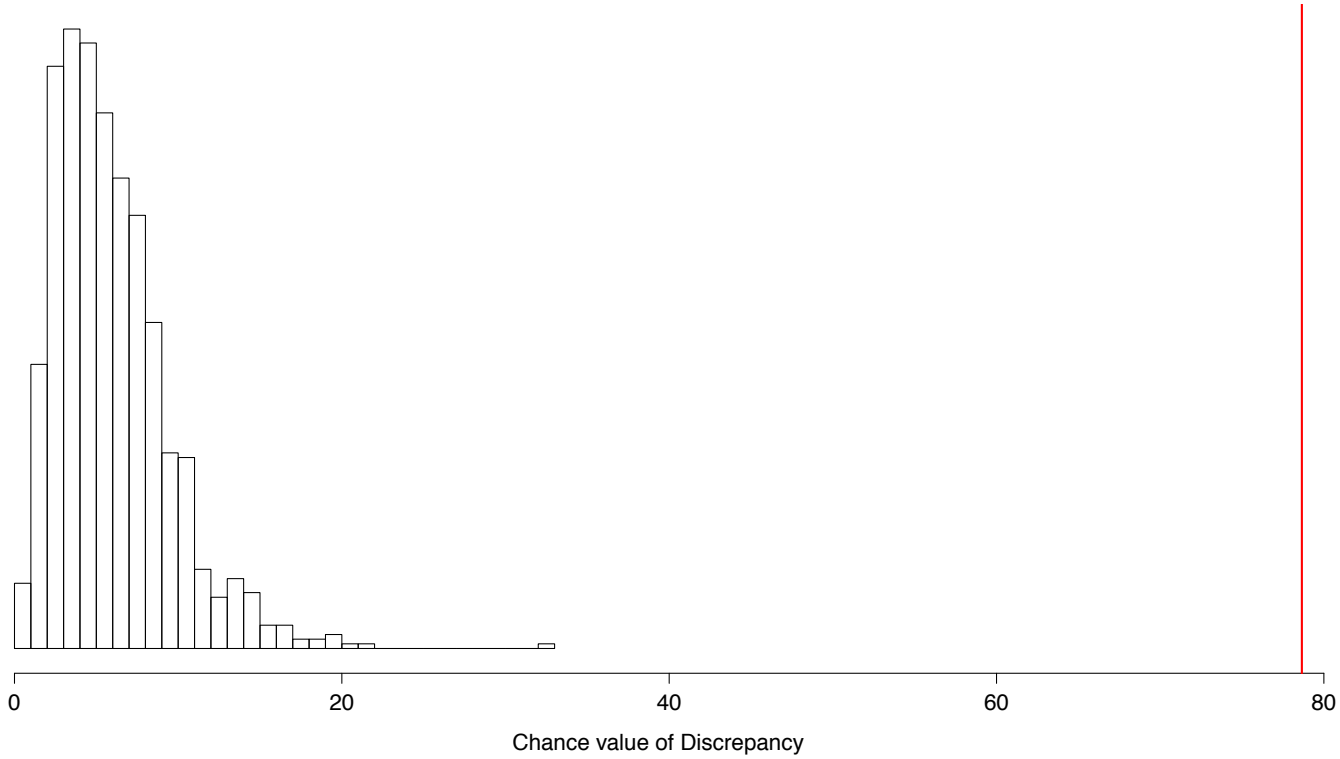
Example: dropped call data output 4

The top plot is the mosaic plot along with a segmented bar chart of the marginal distribution of y.



Example: dropped call data output 5

The bottom plot is the distribution of the values of discrepancy calculated on the permutation datasets (where x and y by design have no association). The red line indicates the value of the discrepancy observed in your data.



Example: dropped call data conclusion

Since the p-value is less than 5%, we conclude that the association between provider and dropped call frequency is statistically significant. The implication is that the distribution of dropped call frequency is somehow different between carriers (i.e., not all carriers are created equal).

Looking at the mosaic plot, the differences in carriers is large and of practical significance. It looks like back in 2009 Knoxville, Verizon was definitely the provider to have. Things may have changed since then.

Statistical vs. Practical significance

Statistical vs. Practical significance

An association is **statistically significant** if it is unlikely that a pair of unrelated variables would exhibit such a large difference in conditional distributions (i.e., at least as large a discrepancy between observed and expected counts) as the ones that we observed in our data.

- This does *not* mean that the association is strong, interesting, or important
- For very large sample sizes, even extremely weak associations can be statistically significant.

Always look at the plots for signs of **practical significance**, i.e., a large enough difference that matters and is meaningful to you (subjective).

Note about *p*-values

Important note about *p*-values of test

The *p*-value is *estimated* from the permutation procedure as the fraction of datasets where D exceeded the observed value by chance. Due to natural variability (you don't expect a fair coin to land heads exactly 250 out of 500 flips), *there is still a little uncertainty in the quoted p-value* (the number may be different if you change the random number seed).

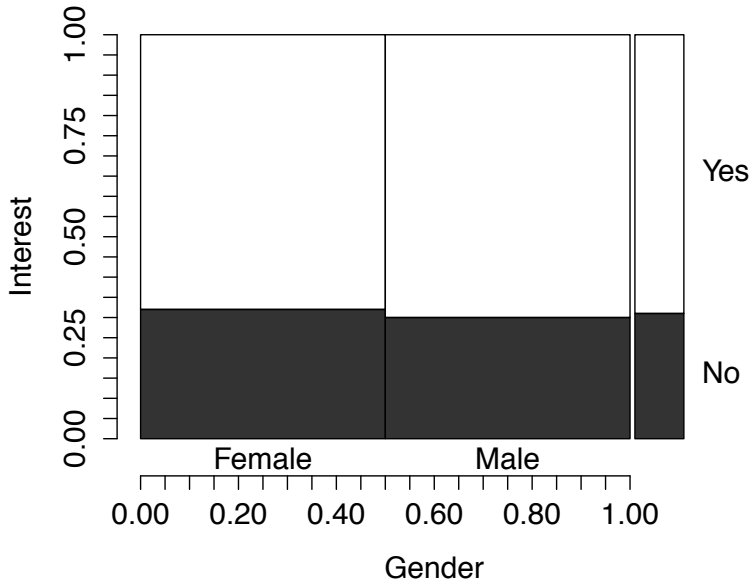
The output in R gives a *range* for the *p*-value of the test. If the 0.05 is inside the range of quoted *p*-values, the test is inconclusive. Increase the number of permutations until you are confident that the *p*-value is either below 5% or above 5%.

Example - frequency flier interest

A survey of 50 people asked whether they would be interested in a new frequent flier program. Very small difference in distributions and p-value says no significant association.

```
data(SMALLFLYER)
associate(Interest~Gender,data=SMALLFLYER)
```

```
                Discrepancy    pvalue
Permutation test  0.04675082 1.0000000
```



Example - frequency flier interest (cont)

Now imagine just duplicating the original dataset 1000 times and rerunning the test – the discrepancy is highly significant even though the difference in distributions still looks VERY small.

```
data(LARGEFLYER)
associate(Interest~Gender,data=LARGEFLYER)
```

```
  Discrepancy Estimated p-value
     46.65736                 0
```

Example - clicking

Earlier we saw that the fraction of users who click on an ad seemed to vary somewhat on the type of device they were using. What about the position of the ad?

```
data(EX6.CLICK)
associate(Click~BannerPosition,data=EX6.CLICK)
```

```
  Discrepancy Estimated p-value
     88.04572                 0
```

Example - clicking

The association is highly statistically significant, so it looks like the position "matters". However, with nearly 14000 observations even very weak associations will be statistically significant. Indeed, the mosaic plot shows that the difference is relatively small, so it doesn't matter "much" (the practical significance is somewhat low, but may be exploitable).