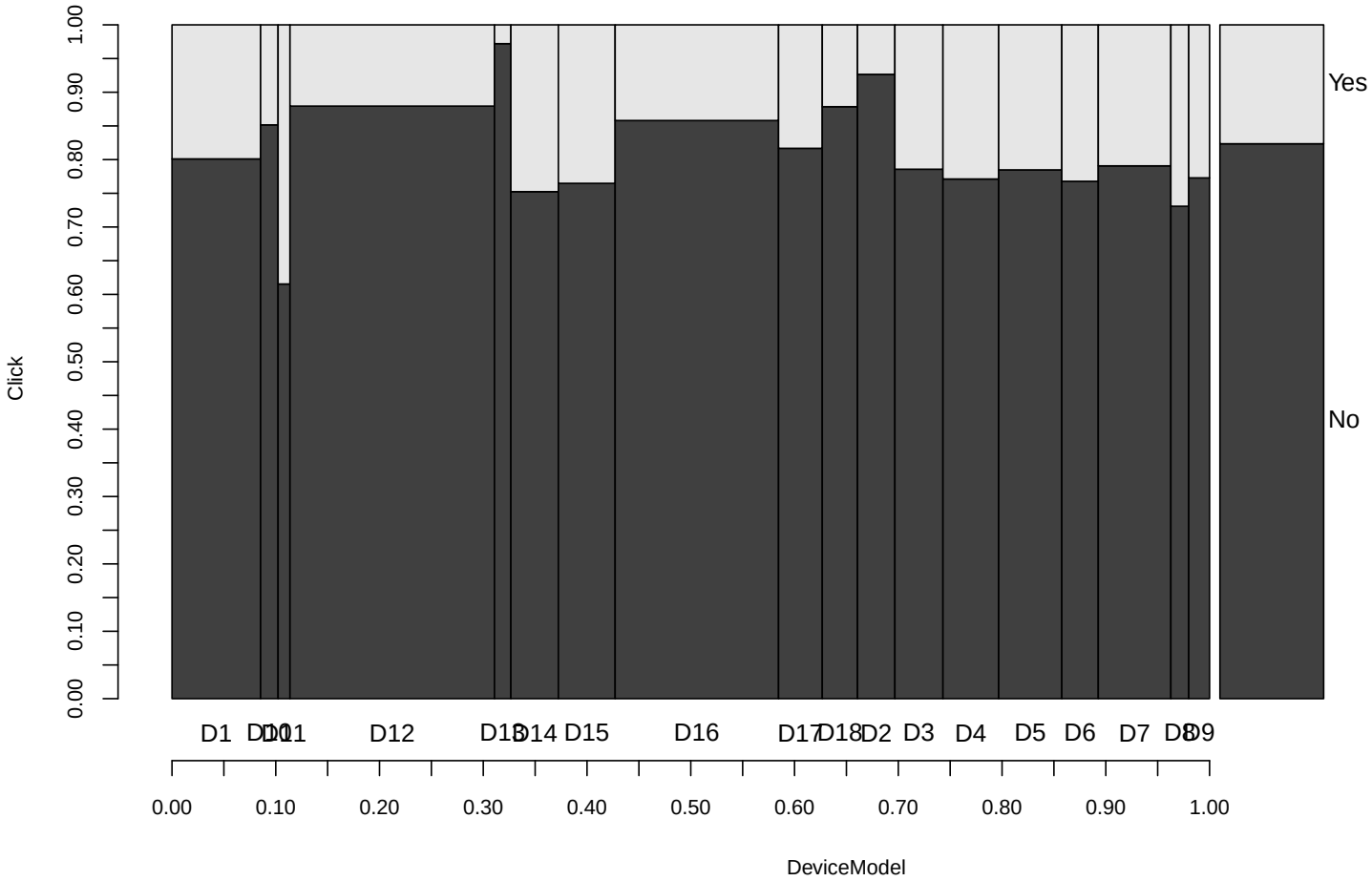
Example - click-thru and device model



Bar widths

The *widths* of each bar tell you the relative fraction of individuals with each level of x.

> When we look for the existence of an association, the widths of the bars are not important because they give information regarding the distribution of x, **not** whether the distribution of y varies between levels of x.

Last slide: the bar for White is the thickest because most of the guys in the survey were white.

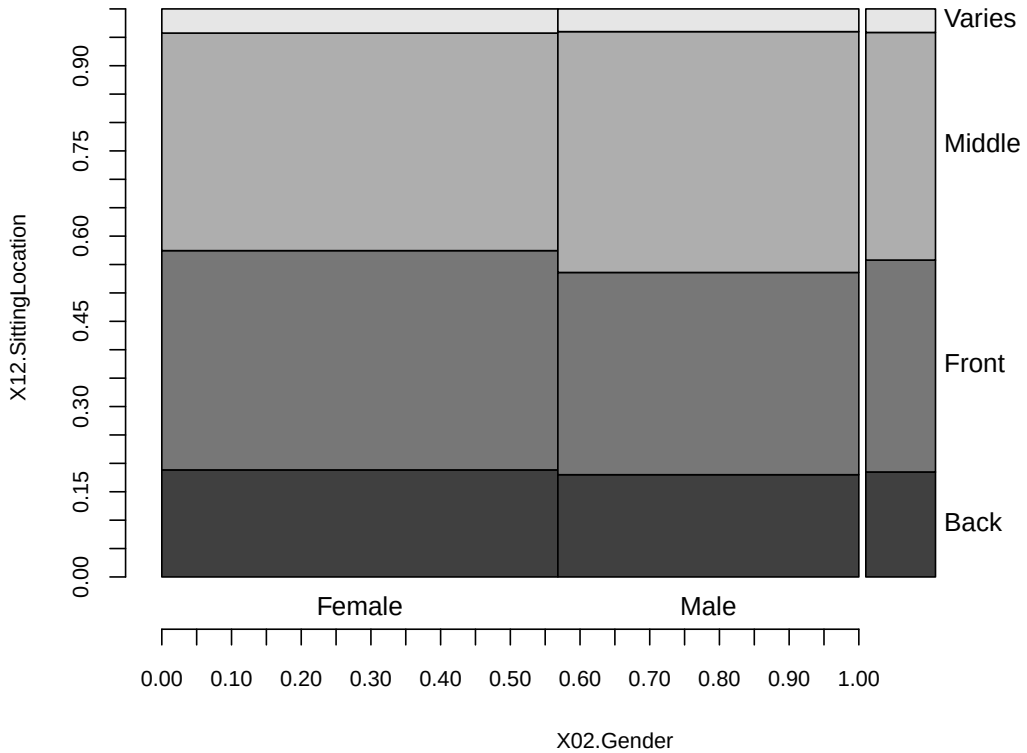
What to look for

The first step in the analysis is to determine whether the association "looks" meaningful, i.e., it is strong enough to have practical importance. To do this, **look to see if the shadings in the bars are more or less the same**.

- If each bar is similar, the distribution of y is about the same for all levels of x. Knowing x doesn't narrow down the possible values of y much, so there is little to no association.
- If there are noticeable differences among bars, the distribution of y differs between at least two levels of x. Knowing x narrows down the possible values of y a lot, so there is an association.

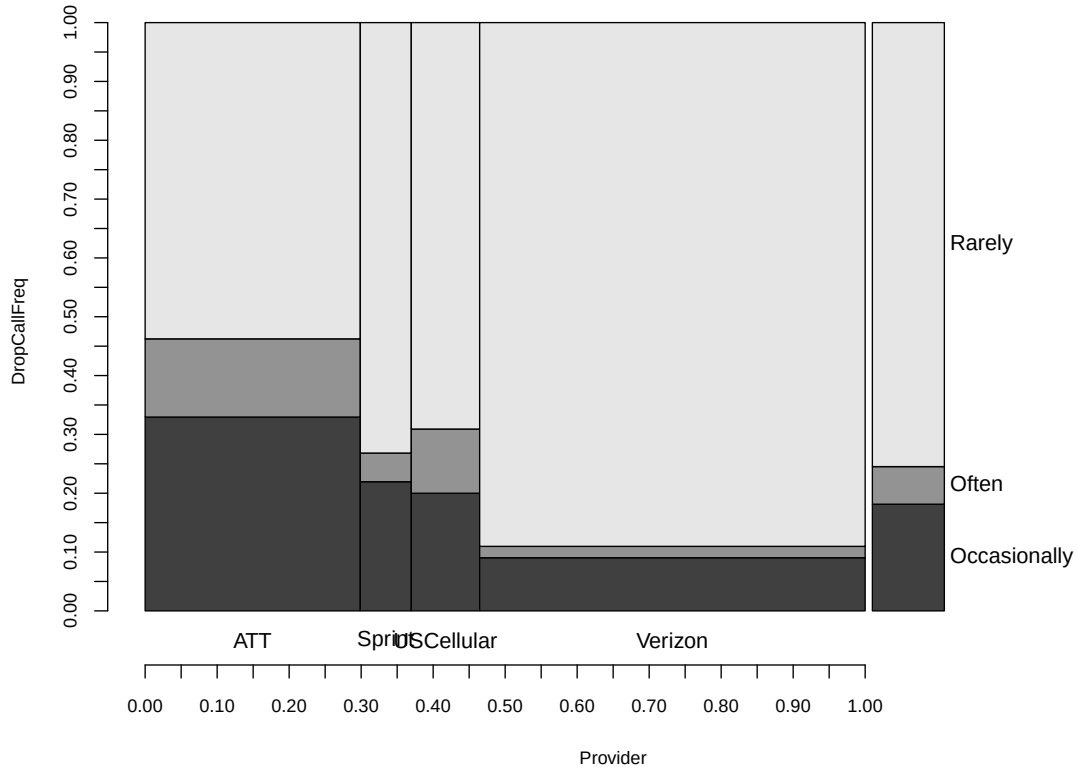
Another example - UT first choice?

```
data(SURVEY09); mosaic(X12.SittingLocation~X02.Gender, data=SURVEY09)
```



The distribution of sitting location is nearly identical for both genders, so we do not suspect an association.

Example - Dropped Calls



The distribution of "Dropped Call Frequency" looks different between providers (e.g., Verizon has a big chunk for "Rarely" while ATT does not).

The bars look *very* different indicating a strong association.

Contingency table

To determine if there is a statistically meaningful association, we need to quantify just **how** the distributions of y vary between levels of x. How should we operationalize the definition of "difference in distributions"?

We start with the **contingency table**. This gives the number of individuals who have each combination of x and y values.

	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

Contingency Table Analysis

The numbers in a contingency table are the **observed counts** in the data.

	Occasionally	Often	Rarely	Total
ATT	57	23	93	173
Sprint	9	2	30	41
USCellular	11	6	38	55
Verizon	28	6	276	310
Total	105	37	437	579

- 38 USCellular customers report Rarely having dropped calls (cell inside table)
- There are a total of 41 Sprint customers (right margin)
- There are 105 people who report Occasionally having dropped calls (bottom margin)
- The sample size is 579 people (bottom right)

Marginal Distribution

The **marginal** (or overall) distribution of a variable is the overall frequency distribution of its levels.

The marginal distribution can be obtained by looking at the numbers in the margins of the contingency table. These give the row and column totals and show the overall (marginal) distribution of x and of y, respectively.

The numbers are often converted into percentages by dividing by the sample size.

Marginal Distribution of Cell Phone Carrier

	Occasionally	Often	Rarely	Total
ATT	57	23	93	**173**
Sprint	9	2	30	**41**
USCellular	11	6	38	**55**
Verizon	28	6	276	**310**
Total	105	37	437	579

The marginal distribution of Cell Phone Carrier is

ATT	Sprint	USCellular	Verizon
173	41	55	310
0.30	0.07	0.09	0.54

The percentages come from taking the count and dividing by the sample size, i.e. $55/549 = 0.09$, after rounding.

Marginal Distribution of Dropped Call Frequency

	Occasionally	Often	Rarely	Total
ATT	57	23	93	173
Sprint	9	2	30	41
USCellular	11	6	38	55
Verizon	28	6	276	310
Total	**105**	**37**	**437**	579

The marginal distribution of Dropped Calls is

Occasionally	Often	Rarely
105	37	437
0.18	0.06	0.75

The percentages are found by taking the counts and dividing by the sample size, i.e. $105/549 = 0.18$, after rounding.

Conditional distribution

The **conditional** distribution of y (given a particular level of x) is the percentages of individuals with that particular level of x who have each level of y. In other words, the conditional distribution of y given x is the relative frequency distribution of y for a **particular** level of x.

When there is no association, the distribution of y for **each** level of x should be similar to the marginal distribution. In other words, knowing x does NOT narrow down the possible values of y.

Conditional Distribution for ATT customers

	Occasionally	Often	Rarely	Total
ATT	57	23	93	173
Sprint	9	2	30	41
USCellular	11	6	38	55
Verizon	28	6	276	310
Total	105	37	437	579

The conditional distribution of dropped call frequency for ATT is

Occasionally	Often	Rarely
57	23	93
0.33	0.13	0.54

The percentages are found by taking the counts and dividing by the number of ATT customers, e.g., $57/173 = 0.33$, after rounding.

Conditional Distributions and the Mosaic Plot

The mosaic plot shows the conditional distribution of y for each level of x.

If there is no association, then the conditional distribution of y for each level of x should closely resemble the marginal (overall) distribution of y.

In other words, when y and x have no association then y has the same distribution regardless of x (no information about x can be leveraged to learn anything about y).

Conditional Distributions for dropped calls

Calculating the conditional distributions of dropped call frequency for each carrier we get:

Carrier	Occasionally	Often	Rarely
ATT	0.33	0.13	0.54
Sprint	0.22	0.05	0.73
USCellular	0.20	0.11	0.69
Verizon	0.09	0.02	0.89
Marginal	0.18	0.06	0.75

There are *substantial* differences between the marginal distribution of y and the conditional distributions of y. Knowing the cell phone carrier greatly narrows down the possibilities for dropped call frequency. We thus suspect an association (this is the same information we gained in the mosaic plot).

Defining discrepancy between distributions

There are many ways operationalize the definition of "difference" in conditional distributions. We will use the popular method of comparing the observed counts (O) in the contingency table to the counts we would have *expected* (E) had their been no association.

We will call the difference in conditional distribution by the word **discrepancy** and define it as:

$$D = \text{sum over every value in table of } \frac{(O-E)^2}{E}$$

O is given from the table. E must be computed.

Expected counts

The formula for the discrepancy between distribution involves the number we would *expect* in the contingency table had there been no association. There is a formula which gives us this number.

For each value in the table, E is:

$$E = \frac{\text{Row Total} \times \text{Column Total}}{n}$$

Finding expected counts

Let's illustrate where this formula comes from with the dropped call data by answering the question:

How many Verizon customers would we expect to Rarely have dropped calls if there was no association between carrier and dropped call frequency?

- If there was no association, then the distribution of dropped call frequency should be the same for each carrier.
- The distribution for Verizon would be the same as the overall, marginal distribution.

Occasionally	Often	Rarely
0.18134715	0.06390328	0.7547496

- Overall, 75% of people rarely experience dropped calls. If there were no association, this should be true for Verizon customers as well. There are 310 Verizon customers, so we expect 75.475% of them, or $0.75475 \times 310 = 234.0$ to report rarely having dropped calls.

Finding expected counts

The expected count was essentially $0.75 \times 310 = 234.0$.

- The 75% can be seen to be the column total for Rarely divided by the sample size (437/579)
- 310 is the row total for this Verizon.
- Putting this together we get the row total times column total divided by the "grand total" (sample size).

$$E = \frac{\text{Row Total} \times \text{Column Total}}{n}$$

Calculating D

Once each expected count has been calculated, the tedious process of calculating D can occur. Software will do this for us.

Observed	Occasionally	Often	Rarely	Total
ATT	57	23	93	173
Sprint	9	2	30	41
USCellular	11	6	38	55
Verizon	28	6	276	310
Total	105	37	437	579

Expected	Occasionally	Often	Rarely
ATT	31.4	11.1	130.6
Sprint	7.4	2.6	30.9
USCellular	10.0	3.5	41.5
Verizon	56.2	19.8	234.0

$$D = \frac{(57-31.4)^2}{31.4} + \frac{(23-11.1)^2}{11.1} + \ldots + \frac{(6-19.8)^2}{19.8} + \frac{(276-234.0)^2}{234.0}$$

$$D = 78.7$$

How big a D is evidence of an association?

The numerical value of D tells us how much of a discrepancy exists between the conditional distributions of y and the marginal distribution of y, and can be calculated from the counts we observed and the counts we'd expect if there is no association.

Values of D farther from zero indicate larger discrepancies and thus stronger associations. The key question is:

> How large does D need to be for us to be convinced that there is an association?

How big a D is evidence of an association?

With the dropped call data, $D = 78.7$. Is this a large value for the discrepancy? Could we have observed this value when all carriers have the same fundamental distribution of dropped call frequency?

We can find the values of D we expect to find "by chance" with a simulation called the **permutation procedure**.