11.0 Contingency Tables

- Discuss Quizzes/Answer Questions
- More Exercises
- Contingency Tables

11.1 More Exercises

Recall the simple version of Bayes' Rule:.

Let $A_1, \ldots, A_n$ be mutually exclusive and suppose that $P[A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n] = 1$. Then

$$P[A_1|B] = \frac{P[B|A_1] * P[A_1]}{\sum_{i=1}^{n} P[B|A_i] * P[A_i]}.$$

This formula is due to the Rev. Thomas Bayes, 1763.

Suppose Karl Rove takes a lie detector test about outing CIA agent Valerie Plame. And suppose a lie detector is 92% accurate on truth tellers, but only 85% accurate on liers.

Also suppose that the leak had to come from one of five people, all of whom are equally suspect.

If Karl Rove denies leaking but the lie detector signals, then what is the probability that he was responsible for the leak?

Set B={signal}, A_1={lie}, A_2={truth}.

$$P[\ \textbf{lie} \mid \textbf{signal}\ ] =$$

$$\frac{P[\ \textbf{signal} \mid \textbf{lie}\ ] * P[\ \textbf{lie}\ ]}{P[\ \textbf{signal} \mid \textbf{lie}\ ] * P[\ \textbf{lie}\ ] + P[\ \textbf{signal} \mid \textbf{truth}\ ] * P[\ \textbf{truth}\ ]}$$

$$\begin{aligned} &= (.85) * (.2)/[(.85) * (.2) + (1 - .92) * (.8)] \\ &= .726 \end{aligned}$$

Thus the guilty signal from the lie detector only increases the probability that Rove was the leak from 1/5 to .726.

11.2 Contingency Tables

A contingency table shows counts for two categorical variables. For example, you might classify a sample of people by gender and major, or by race and marital status, or by diet and health.

	Male	Female
Math	10	20
English	20	10
History	15	15

The Physician's Health Study (1989) was a randomized, controlled, double-blind experiment. Half of 22071 men over 40 took an aspirin every other day, while half took a placebo.

	Heart Attack	No Heart Attack
aspirin	104	10933
placebo	189	10845

The question is whether membership in the aspirin group is independent of whether one is in the heart attack group, or whether aspirin affects (lowers) the heart attack rate.

One way to assess whether or not treatment is independent of heart attack is to make a mosaic plot. In a mosaic plot, the size of the tiles is proportional to the count in the cell. If it looks like a window, one has perfect independence.

.

But no data set is really going to give you a perfect windowpane. Even if the two categories are independent, random chance will cause some variation.

For example, if you toss a coin to determine who gets a heart attack and who does not, then aspirin use is clearly unrelated to heart attack.

But we know that you are unlikely to get exactly equal numbers of heads for both groups. By chance, one or the other will have a slight excess of heart attacks.

We need some way to measure how unlikely the observed data are, if in fact aspirin use has nothing to do with heart attacks.

Determining whether or not an observed contingency table shows significant difference from independence is something we shall study later.

For now, we shall describe two ways of measuring the how far from independence a particular 2×2 contingency table is.

- Relative Risk
- Odds Ratio

To define these measures, consider the following 2×2 contingency table:

	Category 1	Category 2
Type 1	A	B
Type 2	C	D

Relative Risk is

$$RR = \frac{A/(A+B)}{C/(C+D)}.$$

This is just the ratio of the proportion in Category 1 in the first row to the proportion in Category 1 in the second row.

The Odds Ratio is

$$RR = \frac{A/B}{C/D}.$$

This is the ratio of the odds of being in the Category 1 in the first row the to odds of being in Category 1 in the second row.

The odds are just the ratio of the proportion in group to the proportion in the other, where the sum of the proportions must add to 1. Thus odds of "2 to 1" mean that the first outcome is twice as likely as the second, or the chance of the first outcome is $2/3$, while the chance of the second is $1/3$.

In the case of aspirin data:

	Heart Attack	No Heart Attack
aspirin	104	10933
placebo	189	10845

$$
\begin{aligned}
RR &= \frac{104/(104+10933)}{189/(189+10845)} = .55 \\
OR &= \frac{104/10933}{189/10845} = .546.
\end{aligned}
$$

It is important to get the “A” cell right—it should be the bad outcome and the treatment of interest.

If the mosaic plot is a perfect windowpane, then both the Relative Risk and the Odds Ratio equal to 1. But if either is much less than 1, that is strong evidence of dependence (i.e., aspirin helps).

If the first row people are more likely to be in the first column than second row people, then both Relative Risk and the Odds Ratio are greater than 1.

If the first row people are less likely to be in the first column than second row people, then both Relative Risk and the Odds Ratio are less than 1.

Thus one can tell whether aspirin hurts, helps, or is independent of heart attack.

As in our example, it often happens that both Relative Risk and the Odds Ratio have similar, but not identical, values.

Statisticians slightly prefer Relative Risk as a measure of dependence, since it is interpretable has the ratio of two probabilities.

For historical reasons, the Odds Ratio is often used in medicine and gambling. But it is slightly more difficult to interpret, at least for me.