**Topic 6**

**Independent Events**

Events (or Outcomes) are independent iff their joint probabilities are a product of their probabilities.



Variables that seem like they should be related may express this behavior.

Usually variables that seem related have association.

**Method 1:**

This equation translates to the following in two way tables: The joint distribution in a 2 way table is independent if the relative frequency of the joint distribution is the product of its marginal frequencies.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always | 880 | 227 | 1107 |
| Never | 128 | 32 | 160 |
| Total | 1008 | 259 | 1267 |

With frequencies

|  |  |  |  |
| --- | --- | --- | --- |
|   | Females | Males | Total |
| Always | 0.69 | 0.18 | 0.87 |
| Never | 0.10 | 0.03 | 0.13 |
| Total | 0.80 | 0.20 | 1.00 |

Note that p\_always x p\_Females = p\_AlwaysFemales, or .87 x .80 = .69 (not exact due to round off err.)

**Method 2:**

Categorical variables are said to be independent if the conditional distributions of the different values of a variable are identical. In a graph of relative frequencies on rows (conditioned on row values)

|  |  |  |  |
| --- | --- | --- | --- |
|   | Females | Males | Total |
| Always | 0.80 | 0.20 | 1 |
| Never | 0.80 | 0.20 | 1 |

In a graph of relative frequencies on columns (conditioned on column values)

|  |  |  |
| --- | --- | --- |
|   | Females | Males |
| Always | 0.87 | 0.87 |
| Never | 0.13 | 0.13 |
| Total | 1.00 | 1.00 |

**RELATIVE RISK**

|  |  |
| --- | --- |
|   | Lung Disease |
| Present | Absent |
| Smoker | X11 | X12 |
| NonSmoker | X21 | X22 |

$$Relative Risk=\frac{x\_{11}/(x\_{11}+x\_{12})}{x\_{21}/(x\_{21}+x\_{22})}= \frac{\hat{p}\_{when exposed}}{\hat{p}\_{when not exposed}}$$

Essentially, the relative risk tells one how much more likely disease will occur if a subject is exposed to a treatment (treatment might be something bad, like chemical exposure, drug use, radiation, or living for years with someone that hums constantly).

|  |  |
| --- | --- |
|   | Lung Disease |
| Present | Absent |
| Smoker | 100 | 50 |
| NonSmoker | 20 | 180 |

Then the relative risk = rr = [100/(100+50)]/[20/(20+180)] = (100/150)/(20/200)= 6.67

**Odds Ratios** are also popular and well known. $Odds Ratio= \frac{x\_{11}/x\_{12 }}{x\_{21}/x\_{22}} $

OR = (100/20)(50/180) = 1.33

Odd ratios are the ratio of odds of a disease in the exposed group to the odds in the unexposed group.

Odd ratios can give wild numbers that are hard to interpret. Log(OR) is often used for that reason.

Comparison of the two measures:

* Relative Risk is used in randomized control trials.
* Odds Ratios are used in case-controlled studies.
* RR $\~$ OR when x11 and x21 are small.

**Simpson’s Paradox**

Simpson paradox occurs when aggregated or combined proportions reverse in the direction of their relationship relative to the disaggregated proportions.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **1995** | **1996** | **Combined** |
| Derek Jeter | 12/48 | .250 | 183/582 | .314 | 195/630 | **.310** |
| David Justice | 104/411 | **.253** | 45/140 | **.321** | 149/551 | .270 |

Simpson’s Paradox also occurs in continuous data. X might be the amount of a drug, with y as the response and the two lines representing men and women. Note that this is not due to an imbalance in the population, but to the differences in baseline response between the two groups.

