People often feel that the important thing about Paradoxes is *why* they occur. Another very important aspect of paradoxes is knowing *when* they might occur, that way one knows when to look for them and is not surprised by them.

So let’s look at the (minimum) set-up for Simpson’s Paradox:

Simpson’s paradox occurs

* randomly in about 1/60 of 2-way tables
* when there is cross imbalance cell counts

|  |  |
| --- | --- |
| BIG n | small n |
| small n | BIG n |

Two-way table set-up:

**2 variables**

 Two Categorical variables

**Proportions are presented in**

A two-way table

**3 variables**

 One quantitative variable

 Two Categorical variables

**Proportions are presented in**

A nested Two-way table (or two two-way tables)

**Simpsons paradox occurs when the variable that splits the two tables is eliminated and the tables are combined**, and the proportions displayed change in value so that **the category with the larger proportions becomes the category with the smaller proportions.**

Note that when we combine a category, we calculate proportions along the rows instead of the columns, so we are calculating numbers based on a different total n.

Example: Two variables

|  |  |  |
| --- | --- | --- |
|  | **Treatment A** | **Treatment B** |
| **Small Stones** | **93% (81/87)** | 87% (234/270) |
| **Large Stones** | **73% (192/263)** | 69% (55/80) |
| **Both** | 78% (273/350) | **83% (289/350)** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1995** | **1996** | **1997** | **Combined** |
| Derek Jeter | 12/48 | .250 | 183/582 | .314 | 190/654 | .291 | 385/1284 | **.300** |
| David Justice | 104/411 | **.253** | 45/140 | **.321** | 163/495 | **.329** | 312/1046 | .298 |

Example: Three variables

**By Victim’s Race**

|  |  |  |
| --- | --- | --- |
| Caucasian |  | Death Penalty |
|  |  | Yes | No |
| Defendant’s Race | Caucasian | 53 (11.3%) | 414 |
| AfricanAmerican | 11 (22.9) | 37 |

|  |  |  |
| --- | --- | --- |
| AfricanAmerican |  | Death Penalty |
|  |  | Yes | No |
| Defendant’s Race | Caucasian | 0 (0%) | 16 |
| AfricanAmerican | 4 (2.8%) | 139 |

**Combined**

|  |  |  |
| --- | --- | --- |
|  |  | Death Penalty |
|  |  | Yes | No |
| Defendant’s Race | Caucasian | 53 (11%) | 430 |
| AfricanAmerican | 15 (7.9%) | 176 |

Why does this occur? When we look at the second set of proportions, we are calculating the proportions against the row total rather than the column total.

The row total is a different number.

Often we only know the ratios or percentages. Since n is part of calculating $\hat{p}$ or %, it can impact the result.

**When the n for categories is highly imbalanced, simpson’s paradox can occur.**

Note that the two tables above can be combined to make a nested table, from which one category can be rolled up.

|  |  |  |
| --- | --- | --- |
| Race |  | Death Penalty |
| Victim | Defendant | Yes | No |
| Caucasian | Caucasian | 53 (11.3%) | 414 |
| AfricanAmerican | 11 (22.9) | 37 |
| AfricanAmerican | Caucasian | 0 (0%) | 16 |
| AfricanAmerican | 4 (2.8%) | 139 |