SIMPSON'S PARADOX

Consider events E, F and G in sample space Ω. Then, by the Theorem of Total Probability, we have

$$\mathrm{P}(E) = \mathrm{P}(E|F)\mathrm{P}(F) + \mathrm{P}(E|F')\mathrm{P}(F').$$

Now, consider the conditional probability $\mathrm{P}(E|G)$. From the definition of conditional probability,

$$\begin{aligned}
\mathrm{P}(E|G) &= \frac{\mathrm{P}(E\cap G)}{\mathrm{P}(G)} = \frac{\mathrm{P}((E\cap F\cap G))\cup(E\cap F'\cap G))}{\mathrm{P}(G)} = \frac{\mathrm{P}(E\cap F\cap G)+\mathrm{P}(E\cap F'\cap G)}{\mathrm{P}(G)} \\
&= \frac{\mathrm{P}(E|F\cap G)\mathrm{P}(F\cap G)+\mathrm{P}(E|F'\cap G)\mathrm{P}(F'\cap G)}{\mathrm{P}(G)} \\
&= \mathrm{P}(E|F\cap G)\mathrm{P}(F|G)+\mathrm{P}(E|F'\cap G)\mathrm{P}(F'|G),
\end{aligned}$$

(assuming that all conditional probabilities are well-defined), so the Theorem of Total Probability can also be applied to conditional probabilities such as $\mathrm{P}(E|G)$.

Thus we have

$$\begin{aligned}
\mathrm{P}(E|G) &= \mathrm{P}(E|F\cap G)\mathrm{P}(F|G)+\mathrm{P}(E|F'\cap G)\mathrm{P}(F'|G) \\
\mathrm{P}(E|G') &= \mathrm{P}(E|F\cap G')\mathrm{P}(F|G')+\mathrm{P}(E|F'\cap G')\mathrm{P}(F'|G')
\end{aligned}$$

It is possible to construct examples in which

$$\mathrm{P}(E|F\cap G) > \mathrm{P}(E|F'\cap G) \;\text{ and }\; \mathrm{P}(E|F\cap G') > \mathrm{P}(E|F'\cap G') \;\text{ but }\; \mathrm{P}(E|G) < \mathrm{P}(E|G')$$

or, conversely, in which

$$\mathrm{P}(E|F\cap G) < \mathrm{P}(E|F'\cap G) \;\text{ and }\; \mathrm{P}(E|F\cap G') < \mathrm{P}(E|F'\cap G') \;\text{ but }\; \mathrm{P}(E|G) > \mathrm{P}(E|G').$$

To see this, consider a partition of Ω into the eight subsets

$$\begin{array}{cccc}
E\cap F\cap G & E\cap F\cap G' & E\cap F'\cap G & E\cap F'\cap G' \\
E'\cap F\cap G & E'\cap F\cap G' & E'\cap F'\cap G & E'\cap F'\cap G'
\end{array}$$

with associated probabilities

$$\begin{array}{cccc}
p_1 & p_2 & p_3 & p_4 \\
p_5 & p_6 & p_7 & p_8
\end{array}$$

where, say, $0 < p_i < 1$, and $\sum_{i=1}^{8} p_i = 1$. Then we have, for example,

$$\mathrm{P}(E|G) = \frac{p_1+p_3}{p_1+p_3+p_5+p_7} \qquad \mathrm{P}(E|G') = \frac{p_2+p_4}{p_2+p_4+p_6+p_8}$$

$$\mathrm{P}(F|G) = \frac{p_1+p_5}{p_1+p_3+p_5+p_7} \qquad \mathrm{P}(F|G') = \frac{p_2+p_6}{p_2+p_4+p_6+p_8}$$

$$\mathrm{P}(E|F\cap G) = \frac{p_1}{p_1+p_5} \qquad \mathrm{P}(E|F'\cap G) = \frac{p_3}{p_3+p_7}$$

$$\mathrm{P}(E|F\cap G') = \frac{p_2}{p_2+p_6} \qquad \mathrm{P}(E|F'\cap G') = \frac{p_4}{p_4+p_8}$$

Hence

$$P(E|F \cap G) > P(E|F' \cap G) \implies p_1p_7 > p_3p_5 \quad \text{and} \quad P(E|F \cap G') > P(E|F' \cap G') \implies p_2p_8 > p_4p_6$$

Clearly, from these inequalities, we **cannot** conclude anything about the comparative magnitudes of $P(E|G)$ and $P(E|G')$. For example, we cannot conclude that

$$P(E|G) > P(E|G') \quad \text{that is} \quad (p_1 + p_3)(p_4 + p_8) > (p_2 + p_6)(p_5 + p_7).$$

Similarly, no such conclusions can be made if the directions of the inequalities are reversed.

Interpretation: The probability results described above are mathematically straightforward, and follow directly from the definition of conditional probability, and the Theorem of Total Probability. However, the results illustrate that great care must be taken when interpreting conditional probabilities.

It is possible to construct practical examples in which the casual interpretation of results of data collection exercises can lead to incorrect conclusions. Such examples illustrate **Simpson's Paradox**.

SIMPSON'S PARADOX: EXAMPLES

EXAMPLE 1 The examination performance of a class of 80 students is to be studied. Each student is asked to select either an essay or a multiple choice format, and pass rates are examined. It is of interest to assess any difference in examination performance between females and males.

Define event G to be that a student is female, and E to be that a student passes the exam. The results for the 80 students can be presented as a 2×2 table

	E	E'
G	20	20
G'	24	16

from which it is concluded that

$$P(E|G) = \frac{20}{20+20} = \frac{1}{2} \qquad P(E|G') = \frac{24}{24+16} = \frac{3}{5} \qquad \implies \qquad P(E|G) < P(E|G')$$

and so the pass rate amongst females is **lower** than the pass rate amongst males.

However, the results for the two exam formats considered separately indicate contradictory conclusions. For the multiple choice paper, the results table is

	E	E'
G	8	2
G'	21	9

from which it is (casually) concluded that

$$P(E|G) = \frac{8}{8+2} = \frac{4}{5} \qquad P(E|G') = \frac{21}{21+9} = \frac{7}{10} \qquad \implies \qquad P(E|G) > P(E|G')$$

and so, on the multiple choice paper, the pass rate amongst females is **higher** than the pass rate amongst males. Similarly, for the essay paper, the results table is

	E	E'
G	12	18
G'	3	7

from which it is (casually) concluded that

$$\mathrm{P}(E|G) = \frac{12}{12+18} = \frac{2}{5} \qquad \mathrm{P}(E|G') = \frac{3}{3+7} = \frac{3}{10} \qquad \Longrightarrow \qquad \mathrm{P}(E|G) > \mathrm{P}(E|G')$$

and so, on the essay paper, the pass rate amongst females is **higher** than the pass rate amongst males.

Hence, it appears that although females perform better than males in both exam formats, the conclusion is that males perform better overall.

This apparent paradox is resolved simply by noting that, the two tables displaying the results for the multiple choice and essay formats actually define probabilities *conditional* on event F, that the student opted for a multiple choice exam, and on F' respectively. That is, we should conclude from these tables that

$$\mathrm{P}(E|F \cap G) = \frac{8}{10} \quad \mathrm{P}(E|F \cap G') = \frac{7}{10} \quad \mathrm{P}(E|F' \cap G) = \frac{4}{10} \quad \mathrm{P}(E|F' \cap G') = \frac{3}{10}$$

It is then clear that, the overall conclusion of a higher male pass rate results due to unequal allocations of females/males to the two tests, that is,

$$\mathrm{P}(F|G) = \frac{10}{40} = \frac{1}{4} \qquad \mathrm{P}(F|G') = \frac{30}{40} = \frac{3}{4}.$$

The higher male pass rate overall is the result of proportionately more males opting for the multiple choice format exam, on which the pass rate was markedly higher than for the essay format exam. Therefore care must be taken when interpreting the "pooled" results.

EXAMPLE 2 An example of the paradox was found with respect to graduate admissions at the University of California at Berkeley. Aggregate data showed a disparity between the percentage of females applicants and the admission rates of females into graduate programs at Berkeley. When the data were broken down by department, the disparity disappeared because of differences in application rates across departments and differences in acceptance rates of all applicants across departments

	Business School			Law School	
	Admit	Deny		Admit	Deny
Male	480	120	Male	10	90
Female	180	20	Female	100	200

It would appear that females are admitted at a slightly higher rate than males in both the Business school and the Law school.

However, if the tables are combined

	Admit	Deny
Male	490	210
Female	280	220

it appears that females are admitted at a lower rate than males.

(see, for example, "Sex Bias in Graduate Admissions: Data from Berkeley", Bickel, P. J., Hammel, P. A., O'Connell, J. W, *Science*, (1975), or "Simpson's Paradox in Real Life", Wagner, C. H., *The American Statistician* (1982))

EXAMPLE 3 In 1972, a one-in-six survey of the electoral roll, largely concerned with thyroid disease and heart disease was carried out in Wickham, a mixed urban and rural district near Newcastle upon Tyne. Twenty years later, a follow-up study was conducted. The results for two age groups of females are shown below; each table shows the twenty-year survival status for smokers and non-smokers.

Age 55-64	Dead	Alive
Smokers	51	64
Non-Smokers	40	81

Age 65-74	Dead	Alive
Smokers	29	7
Non-Smokers	101	28

It appears that a higher percentage of smokers die than non-smokers in each table. However, when the tables are combined

	Dead	Alive
Smokers	80	71
Non-Smokers	141	109

it appears that smokers have a lower death rate. In fact, most of the smokers have died off before reaching the older age classes, and so the higher number of deaths (in absolute numbers) for the non-smokers in the older age classes has obscured the result.

Reference: "Ignoring a covariate: An example of Simpson's Paradox", Appleton, D.R. French, J.M. and Vanderpump, M.P (1996), *The American Statistician*, 50, pp340-341.