Chapter 2 Categorical Data Analysis

Section 2.1: Rates

Categorical Variables

1. **Ordinal**: Categories with natural ordering (e.g., stone size: small/large).
2. **Nominal**: Categories without ordering (e.g., gender: male/female).

Rate Definition

- Rate measures the proportion or percentage for a category:

$$\text{rate}(\mathrm{X}) = \frac{\text{Count of X}}{\text{Total Count}}$$

- Properties:
  - $0 \leq \text{rate}(\mathrm{X}) \leq 1$ (as a fraction)
  - $0\% \leq \text{rate}(\mathrm{X}) \leq 100\%$ (as a percentage)
  - Rates provide a normalized comparison for categorical data.

Example (Treatment Outcomes)

- Success rate:

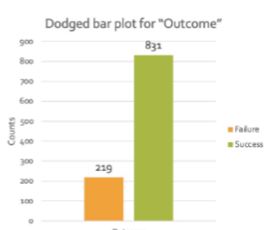
$$\text{rate}(\text{Success}) = \frac{831}{1050} = 0.791\,(79.1\%)$$

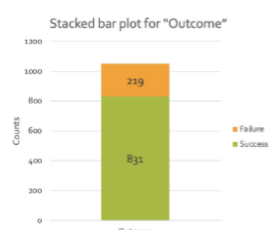
- Failure rate:

$$\text{rate}(\text{Failure}) = \frac{219}{1050} = 0.209\,(20.9\%)$$

Plots

- **Dodged Bar Plot**: Side-by-side comparison of categories.
- **Stacked Bar Plot**: Combined view showing percentages or proportions.





Example

Treatment\Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	*61*	350
Total	831	219	1050

- What proportion of patients *were given treatment Y and had an unsuccessful outcome*?
  - Refers to the joint rate/proportion/percentage.

$$rate(\text{Unsuccessful and Y}) = \frac{61}{1050} = 0.0581$$

- What proportion of patients *given treatment Y had an unsuccessful outcome*?
  - Refers to the conditional rate/proportion/percentage.

$$rate(\text{Unsuccessful} \mid Y) = \frac{61}{350} = 0.174$$

Section 2.2: Association

Definition of Association

- Two variables are **associated** if the presence/absence of one variable changes the rate of another.
  - Positive association: $\text{rate}(\mathrm{A} \mid \mathrm{B}) > \text{rate}(\mathrm{A} \mid \mathrm{NB})$.
  - Negative association: $\text{rate}(\mathrm{A} \mid \mathrm{B}) < \text{rate}(\mathrm{A} \mid \mathrm{NB})$.
  - Note: NB means Not B

Positive association between A and B	Negative association between A and B
$\text{rate}(\mathrm{A} \mid \mathrm{B}) > \text{rate}(\mathrm{A} \mid \mathrm{NB})$	$\text{rate}(\mathrm{A} \mid \mathrm{B}) < \text{rate}(\mathrm{A} \mid \mathrm{NB})$
$\text{rate}(\mathrm{B} \mid \mathrm{A}) > \text{rate}(\mathrm{B} \mid \mathrm{NA})$	$\text{rate}(\mathrm{B} \mid \mathrm{A}) < \text{rate}(\mathrm{B} \mid \mathrm{NA})$
$\text{rate}(\mathrm{NA} \mid \mathrm{NB}) > \text{rate}(\mathrm{NA} \mid \mathrm{B})$	$\text{rate}(\mathrm{NA} \mid \mathrm{NB}) < \text{rate}(\mathrm{NA} \mid \mathrm{B})$
$\text{rate}(\mathrm{NB} \mid \mathrm{NA}) > \text{rate}(\mathrm{NB} \mid \mathrm{A})$	$\text{rate}(\mathrm{NB} \mid \mathrm{NA}) < \text{rate}(\mathrm{NB} \mid \mathrm{A})$

Example

Based on the previous data

> **Example**

Treatment\Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	*61*	350
Total	831	219	1050

- What proportion of patients *were given treatment Y and had an unsuccessful outcome*?
  - Refers to the joint rate/proportion/percentage.

$$rate(\text{Unsuccessful and Y}) = \frac{61}{1050} = 0.0581$$

- What proportion of patients *given treatment Y had an unsuccessful outcome*?
  - Refers to the conditional rate/proportion/percentage.

$$rate(\text{Unsuccessful} \mid Y) = \frac{61}{350} = 0.174$$

- Treatment X success rate:

$$\text{rate}(\text{A} \mid \text{B}) = \text{rate}(\text{Success} \mid \text{X}) = \frac{542}{700} = 0.774\,(77.4\%)$$

- Treatment Y success rate:

$$\text{rate}(\text{A} \mid \text{NB}) = \text{rate}(\text{Success} \mid \text{Y}) = \frac{289}{350} = 0.826\,(82.6\%)$$

- Since

$$\text{rate}(\text{A} \mid \text{B}) < \text{rate}(\text{A} \mid \text{NB})$$

- The success of treatment is **negatively associated with treatment X** because the presence of A when B is present is weaker compared to when B is absent, indicating fewer successful treatments under treatment X.
- Conversely, the success of treatment is **positively associated with treatment Y** because it shows more successful treatments compared to treatment X.

Section 2.3: Two Rules on Rates

Symmetry Rule

1. $\text{rate}(\text{A} \mid \text{B}) > \text{rate}(\text{A} \mid \text{NB}) \iff \text{rate}(\text{B} \mid \text{A}) > \text{rate}(\text{B} \mid \text{NA})$
2. $\text{rate}(\text{A} \mid \text{B}) < \text{rate}(\text{A} \mid \text{NB}) \iff \text{rate}(\text{B} \mid \text{A}) < \text{rate}(\text{B} \mid \text{NA})$
3. $\text{rate}(\text{A} \mid \text{B}) = \text{rate}(\text{A} \mid \text{NB}) \iff \text{rate}(\text{B} \mid \text{A}) = \text{rate}(\text{B} \mid \text{NA})$

	B	Not B	Row Total
A	w	x	$w+x$
Not A	y	z	$y+z$
Column Total	$w+y$	$x+z$	$w+x+y+z$

Based on the first rule

$$\begin{aligned} \text{rate}(\text{A} \mid \text{B}) > \text{rate}(\text{A} \mid \text{NB}) &\iff \text{rate}(\text{B} \mid \text{A}) > \text{rate}(\text{B} \mid \text{NA}) \\ \frac{w}{w+y} > \frac{x}{x+z} &\iff \frac{w}{w+x} > \frac{y}{y+z} \\ w(x+z) > x(w+y) &\iff w(y+z) > y(w+x) \\ wx + wz > xw + xy &\iff wy + wz > yw + yx \\ wz > xy &\iff wz > xy \end{aligned}$$

Example

Based on the previous data

Example

Treatment\Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	*61*	350
Total	831	219	1050

- What proportion of patients *were given treatment Y and had an unsuccessful outcome*?
  - Refers to the joint rate/proportion/percentage.

$$rate(\text{Unsuccessful and Y}) = \frac{61}{1050} = 0.0581$$

- What proportion of patients *given treatment Y had an unsuccessful outcome*?
  - Refers to the conditional rate/proportion/percentage.

$$rate(\text{Unsuccessful} \mid Y) = \frac{61}{350} = 0.174$$

$\text{rate}(\text{A} \mid \text{B}) = \text{rate}(\text{Success} \mid \text{X}) = 0.774$, $\text{rate}(\text{A} \mid \text{NB}) = \text{rate}(\text{Success} \mid \text{Y}) = 0.826$
Using symmetry:
$\text{rate}(\text{B} \mid \text{A}) = \text{rate}(\text{X} \mid \text{Success}) = 0.652 < \text{rate}(\text{B} \mid \text{NA}) = \text{rate}(\text{X} \mid \text{Failure}) = 0.721$

Basic Rule

- Overall rate (A) is will always lie between $\mathrm{rate(A \mid B)}$ and $\mathrm{rate(A \mid NB)}$:

$$\mathrm{rate(A \mid B)} \leq \mathrm{rate(A)} \leq \mathrm{rate(A \mid NB)}$$

Consequence 1:

- The closer $rate(B)$ is to 100%, the closer $rate(A)$ is to $\mathrm{rate(A \mid B)}$.

Consequence 2:

- If $rate(B)$ = 50%, then $\mathrm{rate(A)} = \frac{1}{2}\left[\mathrm{rate(A \mid B)} + \mathrm{rate(A \mid NB)}\right]$

Consequence 3:

- If $\mathrm{rate(A \mid B)} = \mathrm{rate(A \mid NB)}$, then $\mathrm{rate(A)} = \mathrm{rate(A \mid B)} = \mathrm{rate(A \mid NB)}$

Example:
If $\mathrm{rate(A \mid B)} = 75\%$, $\mathrm{rate(A \mid NB)} = 55\%$, overall rate $\mathrm{rate(A)} \in [55\%, 75\%]$.

Section 2.4: Simpson's Paradox

Definition

- A phenomenon where a trend reverses when data is combined vs. when divided into subgroups.
- Simpson's Paradox highlights the importance of subgroup analysis.

Example

Based on the previous data

> ### Example
>
> | Treatment\Outcome | Success | Failure | Row Total |
> | --- | --- | --- | --- |
> | X | 542 | 158 | 700 |
> | Y | 289 | *61* | 350 |
> | Total | 831 | 219 | 1050 |
>
> - What proportion of patients *were given treatment Y and had an unsuccessful outcome*?
>   - Refers to the joint rate/proportion/percentage.
>
> $$rate(\text{Unsuccessful and Y}) = \frac{61}{1050} = 0.0581$$
>
> - What proportion of patients *given treatment Y had an unsuccessful outcome*?
>   - Refers to the conditional rate/proportion/percentage.
>
> $$rate(\text{Unsuccessful} \mid Y) = \frac{61}{350} = 0.174$$
>
> ---

- Combined: $\mathrm{rate(Success \mid X)} < \mathrm{rate(Success \mid Y)}$.
- Split by stone size:
  - Large stones: $\mathrm{rate(Success \mid X)} > \mathrm{rate(Success \mid Y)}$.
  - Small stones: $\mathrm{rate(Success \mid X)} > \mathrm{rate(Success \mid Y)}$.

Section 2.5: Confounders

Definition

A confounder is a third variable associated with both the independent and dependent variables, affecting the observed relationship. Association does not imply causation; consider confounders.

Example

Based on the previous data

> ### Example
>
> | Treatment\Outcome | Success | Failure | Row Total |
> | --- | --- | --- | --- |
> | X | 542 | 158 | 700 |
> | Y | 289 | *61* | 350 |
> | Total | 831 | 219 | 1050 |
>
> - What proportion of patients *were given treatment Y and had an unsuccessful outcome*?
>   - Refers to the joint rate/proportion/percentage.
>
> $$rate(\text{Unsuccessful and Y}) = \frac{61}{1050} = 0.0581$$
>
> - What proportion of patients *given treatment Y had an unsuccessful outcome*?
>   - Refers to the conditional rate/proportion/percentage.
>
> $$rate(\text{Unsuccessful} \mid Y) = \frac{61}{350} = 0.174$$

- **Stone size** is a confounder:
  - Large stones are more likely treated with X ($\text{rate}(\text{Large} \mid \text{X}) = 75.1\%$).
  - Large stones have lower success rates overall ($\text{rate}(\text{Success} \mid \text{Large}) = 71.9\%$).

Addressing Confounders

1. **Slicing**: Analyze subgroups (e.g., by stone size).
2. **Randomization**: Randomly assign patients to treatments to equalize confounders.