

Frage:

Mir wurde heute vom Tutor gesagt, dass $X|Y$ so zu verstehen ist wie: Y gegeben X. Dies bringt mich aber zur Verwirrung. Folgendes das Beispiel im Skript Kapitel 6: (relative Häufigkeiten der bedingten Verteilungen) Fall 1, Geschlecht gegeben Studienfach, also X gegeben Y. Es gilt, die Summe der relativen Häufigkeiten der jeweiligen Studienfächer beträgt 1. Falls , Studienfach gegeben Geschlecht, also Y gegeben X. Es gilt, die Summe der relativen Häufigkeiten der beiden Geschlechter beträgt 1. Analog dazu müsste die Kontingenztafel in Aufgabe 2a laut Aufgabenstellung $X|Y$, also y(Klasse) gegeben X(Überlebensstatus), so aussehen, dass die Summe der relativen Häufigkeiten der geretteten und vermissten Gruppe jeweils 1 beträgt. Dies ist jedoch nicht der Fall, stattdessen beträgt in unserer Tabelle die Summe der relativen Häufigkeiten der jeweiligen Klasse 1. Daher meine Frage, ob ich vielleicht etwas falsch verstanden habe.

Antwort:

Es liegt hier wohl nur eine Verwechslung von X und Y vor. $X|Y$ bedeutet "X gegeben Y", nicht umgekehrt. In Aufgabe Blatt 6/2 werden die Klassen mit Y und "Gerettet/Vermisst" mit X bezeichnet, erkennbar aus der gegebenen Tabelle. Gesucht ist $X|Y$, also "Gerettet/Vermisst" gegeben Klasse (also umgekehrt als in der Frage formuliert). Daher ist es also auch korrekt, dass sich die bedingten relativen Häufigkeiten bezüglich der Klassen zu 1 aufsummieren. Die folgenden Ausführungen sollen allgemein nochmal die Begriffe der gemeinsamen Verteilung, Randverteilung und bedingte Verteilung. Sie beziehen sich auf das Beispiel aus dem Skript Kapitel 6, Seite 3 ff

Gemeinsame Verteilung:

Die gemeinsame Verteilung gibt an, wie die Merkmale Geschlecht (X) und Studienfach (Y) gemeinsam verteilt sind, z.B. 10 Prozent aller Studenten sind männlich und studieren BWL. Die Summe der einzelnen relativen Häufigkeiten $f_{ij} = \frac{n_{ij}}{n}$ ergibt hierbei 1.

Aus der Tabelle der gemeinsamen Verteilung des Geschlechts und Studienfach:

$$0,1 + 0,25 + 0,1 + 0,2 + 0,1 + 0,25 = 1$$

Randverteilung:

Die Randverteilung gibt die Verteilung eines Merkmals, z.B. Geschlecht an, unabhängig vom Studienfach.

$$f_{i+} = \frac{n_{i+}}{n} = \sum_{j=1}^l f_{ij}$$

Die relativen Häufigkeiten einer Zeile werden summiert.

Die Randverteilung für die männlichen Studenten:

$$f_{männlich+} = 0,1 + 0,25 + 0,1 = 0,45$$

Die Randverteilung für die weiblichen Studenten:

$$f_{weiblich+} = 0,2 + 0,1 + 0,25 = 0,55$$

Interpretation:

In der Population der Studenten sind 45% männlich, und 55% weiblich.
Die Summe der Randverteilungen des Geschlechts $f_{Geschlecht+}$ muss 1 ergeben.

$$f_{männlich+} + f_{weiblich+} = 0,45 + 0,55 = 1$$

Die Randverteilung des Studienfachs:

$$f_{+j} = \frac{n_{+j}}{n} = \sum_{i=1}^k f_{ij}$$

Die relativen Häufigkeiten einer Spalte werden summiert. Die Randverteilung des Studienfachs:

$$f_{+BWL} = 0,1 + 0,2 = 0,3$$

$$f_{+VWL} = 0,25 + 0,1 = 0,35$$

$$f_{+Sonstige} = 0,1 + 0,25 = 0,35$$

Interpretation:

In der Population der Studenten studieren 30% BWL, 35% VWL und die restlichen 35% der Studierenden sind in sonstige Studiengänge immatrikuliert.
Die Summe der Randverteilungen des Studienfachs $f_{+Studienfach}$ muss 1 ergeben.

$$f_{+BWL} + f_{+VWL} + f_{+Sonstige} = 0,3 + 0,35 + 0,35 = 1$$

Bedingte Verteilung:

Bei bedingten Verteilungen ist man an der Verteilung eines Merkmals, z.B. Geschlecht (X), gegeben das andere Merkmal (hier Studienfach (Y)) ist festgehalten, interessiert.

Beispielsweise, welcher Anteil der BWL-Studenten ("gegeben eine Person studiert BWL") ist männlich, welcher Anteil ist weiblich?

Aus der Tabelle der gemeinsamen Verteilung, berechnet sich eine bedingte Verteilung $X|Y$ wie folgt:

$$f_{i|j} = \frac{n_{ij}}{n_{+j}} = \frac{f_{ij}}{f_{+j}} = \frac{f_{ij}}{\sum_{i=1}^k f_{ij}}$$

Bedingte Verteilung für der Student ist männlich gegeben er studiert BWL:

$$f_{männlich|BWL} = \frac{0,1}{0,1 + 0,2} = 0,33$$

Bedingte Verteilung für der Student ist weiblich gegeben sie studiert BWL:

$$f_{weiblich|BWL} = \frac{0,2}{0,1 + 0,2} = 0,67$$

Interpretation:

In der Population der BWL-Studierenden sind 67% weibliche und 33% männliche Studenten.

Die Summe der bedingten relativen Häufigkeiten (hier Spaltensumme weil auf eine Spalte bedingt wurde) ergibt 1.

$$f_{männlich|BWL} + f_{weiblich|BWL} = 0,33 + 0,67 = 1$$

Dies gilt für jedes Studienfach.

Die bedingte Verteilung Studienfach (Y) gegeben dem Geschlecht (X):

$$f_{j|i} = \frac{n_{ij}}{n_{i+}} = \frac{f_{ij}}{f_{i+}} = \frac{f_{ij}}{\sum_{j=1}^l f_{ij}}$$

Beispielsweise für alle männlichen Studenten:

$$f_{BWL|männlich} = \frac{0,1}{0,1 + 0,25 + 0,1} = 0,22$$

$$f_{VWL|männlich} = \frac{0,25}{0,1 + 0,25 + 0,1} = 0,56$$

$$f_{Sonstige|männlich} = \frac{0,1}{0,1 + 0,25 + 0,1} = 0,22$$

Die Summe der bedingten relativen Häufigkeiten (hier Zeilensumme) ergibt 1.

$$f_{BWL|männlich} + f_{VWL|männlich} + f_{Sonstige|männlich} = 1$$

Selbes gilt für die bedingte Verteilung Studienfach gegeben weibliche Studenten.