5.3 – Qualitative Methods

Categorical (or qualitative) variables are one whose levels are measured on a nominal scale.

Examples include:

- the mother tongue of an individual,
- her hair colour,
- her age,
- and so forth.

The **central tendency** of the values of such a variable is its **mode**.

Measures of **spread** are much more difficult to define in a consistent manner. One possibility: use the proportion of levels with more than a certain percentage of the observations above a given threshold.

Example: consider a dataset with $n = 517$ individuals and $p = 3$ features:

- **age** $(25-$, $25-44$, $45-64$, $65+)$;
- **mother tongue** (French, English, Mandarin, Arabic, Other), and
- **hair colour** (black, brown, blond, red).

Their respective modes are $25 - 44$, English, and brown. And their spread?

Age	Mother Tongue	Hair Colour			
		Black	Brown	Blond	Red
24-	French	11	24	12	2
	English	12	44	3	6
	Mandarin	16	2	1	0
	Arabic	9	1	0	0
	Other	11	7	13	1
25-44	French	15	32	17	2
	English	21	47	8	7
	Mandarin	23	3	1	0
	Arabic	15	2	0	2
	Other	15	16	12	6
45-64	French	7	12	2	1
	English	4	17	2	3
	Mandarin	3	1	0	0
	Arabic	3	1	0	0
	Other	6	2	1	1
65+	French	1	2	1	0
	English	3	3	0	2
	Mandarin	4	0	0	0
	Arabic	3	0	0	0
	Other	5	1	6	1

Mother Tongue	Hair Colour			
	Black	Brown	Blond	Red
French	34	70	32	5
English	40	111	13	18
Mandarin	46	6	2	0
Arabic	30	4	0	2
Other	37	26	32	9

Age	Hair Colour			
	Black	Brown	Blond	Red
24-	59	78	29	9
25-44	89	100	38	17
45-64	23	33	5	5
65+	16	6	7	3

Mother Tongue	Age			
	24-	25-44	45-64	65+
French	49	66	22	4
English	65	83	26	8
Mandarin	19	27	4	4
Arabic	10	19	4	3
Other	32	49	10	13

Hair Colour			
Black	Brown	Blond	Red
187	217	79	34
36.2%	42.0%	15.3%	6.6%

Mother Tongue				
French	English	Mandarin	Arabic	Other
141	182	54	36	104
27.3%	35.2%	10.4%	7.0%	20.1%

Age			
24-	25-44	45-64	65+
175	244	66	32
33.8%	47.2%	12.8%	6.2%

Total Number of Observations: 517

Percentage of Levels Above:	15%	25%
Hair Colour	75%	50%
Mother Tongue	60%	60%
Age	50%	50%

Qualitative features are often associated to numerical values: in `R`, for instance, there is a difference between factor **levels** and factor **labels**.

Categorical variables with numerical levels are treated as **ordinal** variables.

⚠ **These should not be interpreted as numerals!**

If we use the code "red" $= 1$, "blond" $= 2$, "brown" $= 3$, and "black" $= 4$ to represent hair colour, we **cannot conclude** that "blond" $>$ "red", even though $2 > 1$, or that "black" $-$ "brown" $=$ "red", even though $4 - 3 = 1$.

A categorical variable that has exactly two levels is a **dichotomous** (binary) variable; a variable with more than two levels is **polytomous**.

Regression on categorical variables $\Longrightarrow$ **multinomial logistic regression**.

Distances (apart from the $0-1$ distance and the related Hamming distance) require numerical inputs.

But representing categorical variables with numerical features can lead to traps (see previous slide).

Anomaly detection methods based on distance or on density are not recommended in the qualitative context (unless the distance function has been modified appropriately).

Another option is to look at combinations of feature levels, but this can be computationally expensive.

5.3.1 – AVF Algorithm

The **Attribute Value Frequency** (AVF) algorithm is a fast and simple way to detect outlying observations in categorical data.

It can be done without having to create or search through various combinations feature levels (which increase the search time).

Intuitively, outlying observations are points which occur relatively infrequently in the (categorical) dataset; an "ideal" anomalous point is one for which **each feature value is extremely anomalous** (or relatively infrequent).

The **rarity** of an attribute level can be measured by summing the number of times the corresponding feature takes that value in the dataset.

Let's say that there are n observations in the dataset: $\{\mathbf{x}_i\}$, $i = 1, \ldots, n$, and that each observation is a collection of m features.

We write

$$\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,\ell}, \cdots, x_{i,m}),$$

where $x_{i,\ell}$ is $\mathbf{x}_i$'s ℓth feature's level.

Example: in the previous example, we may have

$$\mathbf{x}_1 = (x_{1,1}, x_{2,1}, x_{3,1}) = (24-, \text{French}, \text{blond})$$

$$\vdots$$

$$\mathbf{x}_{517} = (x_{517,1}, x_{517,1}, x_{517,1}) = (24-, \text{Mandarin}, \text{blond}).$$

The **AVF score** of an observation $\mathbf{x}_i$ is

$$\text{AVFscore}(\mathbf{x}_i) = \frac{1}{m}\sum_{\ell=1}^{m} f(x_{i,\ell}),$$

where $f(x_{i,\ell})$ is the number of dataset observations $\mathbf{x}$ for which the ℓth feature takes on the level $x_{i,\ell}$.

A **low** AVF score indicates that the observation is likely to be an **outlier**.

An "ideal" anomalous observation minimizes the AVF score – reached when the observation's features' levels occurs only once in the dataset.

For an integer k, the suggested outliers are the k observations with smallest AVF scores. The formal procedure is provided in Algorithm 5.

Algorithm 5: AVF

Inputs: dataset D (n observations, m features), number of anomalous observations k

while $i \leq n$ **do**

    $\mathrm{j} = 1$

    $\mathrm{AVFscore}(\mathbf{x}_i) = f(x_{i,j})$

    while $j \leq m$ **do**

        $\mathrm{AVFscore}(\mathbf{x}_i) = \mathrm{AVFscore}(\mathbf{x}_i) + f(x_{i,j})$;

        $j = j + 1$

    end

    $\mathrm{AVFscore}(\mathbf{x}_i) = \mathrm{Mean}(\mathrm{AVFscore}(\mathbf{x}_i))$

    $i = i + 1$

end

Outputs: k observations with smallest AVF scores

Age	Mother Tongue	Hair Colour			
		Black	Brown	Blond	Red
24-	French	167.7	177.7	131.7	116.7
	English	181.3	191.3	145.3	130.3
	Mandarin	138.7	148.7	102.7	87.7
	Arabic	132.7	142.7	96.7	81.7
	Other	155.3	165.3	119.3	104.3
25-44	French	190.7	200.7	154.7	139.7
	English	204.3	214.3	168.3	153.3
	Mandarin	161.7	171.7	125.7	110.7
	Arabic	155.7	165.7	119.7	104.7
	Other	178.3	188.3	142.3	127.3

Age	Mother Tongue	Hair Colour			
		Black	Brown	Blond	Red
45-64	French	131.3	141.3	95.3	80.3
	English	145.0	155.0	109.0	94.0
	Mandarin	102.3	112.3	66.3	51.3
	Arabic	96.3	106.3	60.3	45.3
	Other	119.0	129.0	83.0	68.0
65+	French	120.0	130.0	84.0	69.0
	English	133.7	143.7	97.7	82.7
	Mandarin	91.0	101.0	55.0	40.0
	Arabic	85.0	95.0	49.0	34.0
	Other	107.7	117.7	71.7	56.7

AVF scores; 10 lowest scores highlighted (in red).

$$\begin{aligned}\text{AVFscore}(24-,\text{French},\text{blond}) &= \tfrac{1}{3}(f(24-)+f(\text{French})+f(\text{blond}))\\ &= \tfrac{1}{3}(175+141+79)=131.7\end{aligned}$$

5.3.2 – Greedy Algorithm

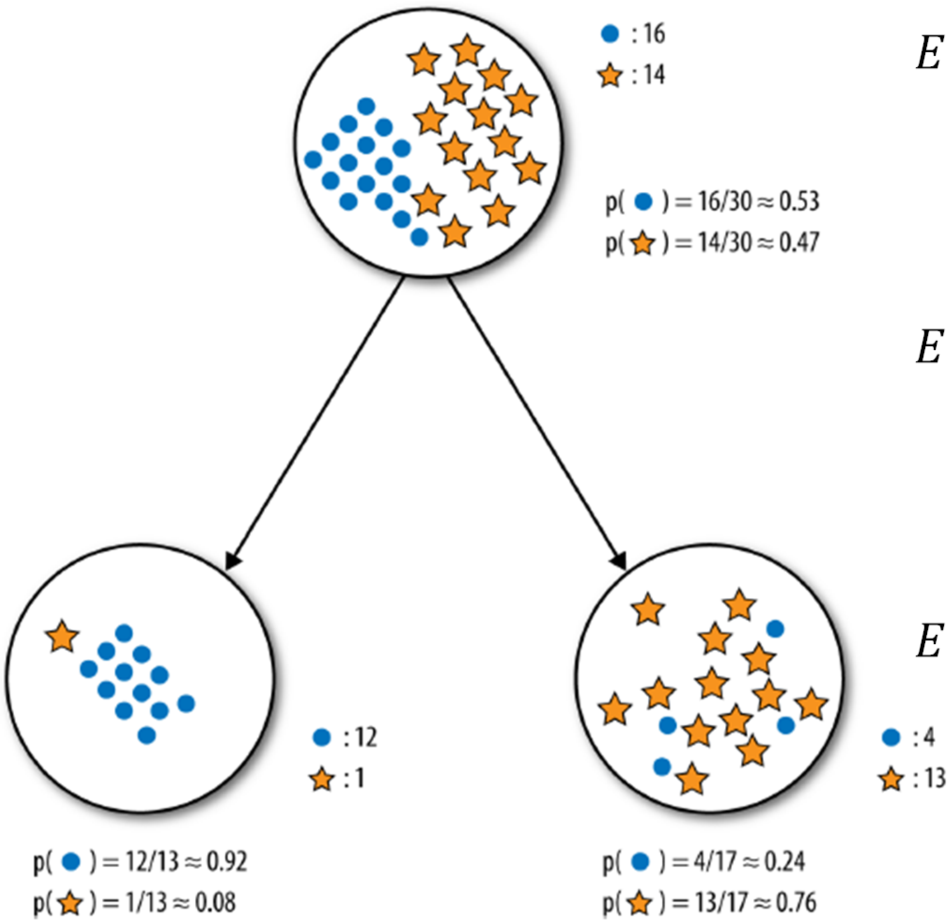
The **greedy** algorithm identifies a set OS of **candidate** anomalous observations in an efficient manner.

The **entropy of a set** $\Sigma \subseteq D$ is a measure of the **disorder** in Σ. Let X be a feature of D; the set of levels that X takes on Σ is denoted by

$$S(X;\Sigma) = \{z | X = z,\ \mathbf{x} \in \Sigma\}.$$

Let $p_X(z)$ be the % of observations in Σ for which $X = z$. The **entropy of a feature** X on Σ is

$$H(X;\Sigma) = -\sum_{z\in S(X;\Sigma)} p_X(z) \log p_X(z).$$



$$
\begin{aligned}
E(S) &= -p_\circ \log p_\circ - p_* \log p_* \\
&= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99
\end{aligned}
$$

$$
\begin{aligned}
E(L) &= -p_\circ \log p_\circ - p_* \log p_* \\
&= -\frac{12}{13}\log\frac{12}{13} - \frac{1}{13}\log\frac{1}{13} \approx 0.39
\end{aligned}
$$

$$
\begin{aligned}
E(R) &= -p_\circ \log p_\circ - p_* \log p_* \\
&= -\frac{4}{17}\log\frac{4}{17} - \frac{13}{17}\log\frac{13}{17} \approx 0.79
\end{aligned}
$$

[Foster, Provost]

The mathematical formulation of the problem is simple: in order to find k anomalous observations in a dataset D, solve the optimization problem

$$\text{OS} = \arg\min_{O \subseteq D} \{H(D \setminus O)\}, \quad \text{subject to } |O| = k,$$

where the **entropy** $H(D \setminus O)$ is the sum of the entropy of each feature:

$$H(D \setminus O) = H(X_1; D \setminus O) + \cdots + H(X_m; D \setminus O)$$
$$H(X_\ell; D \setminus O) = -\sum_{z_\ell \in S(X_\ell; D \setminus O)} p(z_\ell) \log p(z_\ell),$$

where $S(X_\ell; D \setminus O)$ is the set of levels that the ℓth feature takes in $D \setminus O$.

The greedy algorithm solves the optimization problem as follows:

1. The set of outlying and/or anomalous observations OS is initially set to be empty, and all observations of $D \setminus \mathsf{OS}$ are identified as normal.

2. Compute $H(D \setminus \mathsf{OS})$.

3. Every normal observation $\mathbf{x}$ is temporarily taken out of $D \setminus \mathsf{OS}$ to create a subset $D'_{\mathbf{x}}$, whose entropy $H(D'_{\mathbf{x}})$ is also computed.

4. The $\mathbf{y}$ which provides the **maximal entropy impact** is added to OS:

$$\mathbf{y} = \arg \min_{\mathbf{x} \in D \setminus \mathsf{OS}} \left\{ H(D \setminus \mathsf{OS}) - H(D'_{\mathbf{x}}) \right\}.$$

5. Repeat steps 2-4 another $k-1$ times to obtain a set OS of k candidate anomalous observations.