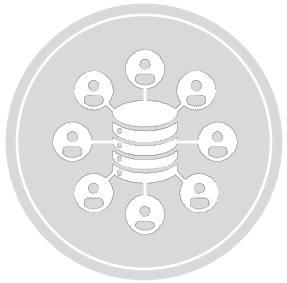
Coding of categorical variables

Study objectives and plan

Experimental design



Data collection

Model estimation

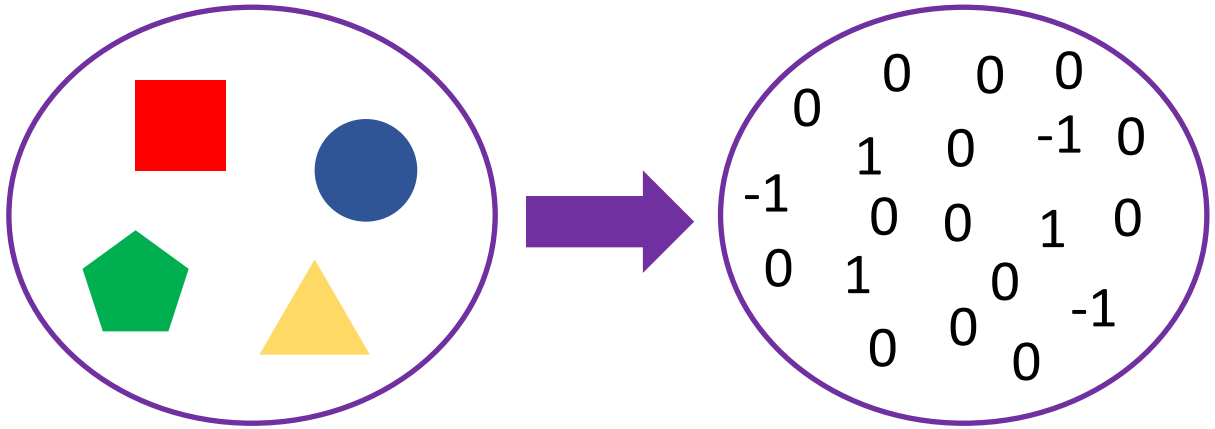
Interpretation

Coding of categorical variables

Numerical coding

- Categorical variables require conversion into numbers when adding to utility functions
- This is done via dummy or effects coding

$$\text{Utility} = \ldots + \alpha \cdot \text{Colour} + \ldots + \beta \cdot \text{SideEffects} + \ldots$$



Coding of categorical variables

Without dummy or effects coding

- Inappropriately presumes a ranking
- Inappropriately presumes equally spaced categories

$$\text{Utility} = \ldots + \alpha \cdot \text{Colour} + \ldots + \beta \cdot \text{SideEffects} + \ldots$$

Colour	Coding
Red	1
Blue	2
Green	3
Yellow	4

Side effects	Coding
Mild	1
Moderate	2
Severe	3

Coding of categorical variables

Dummy coding

- Assume *L* levels (categories), select one level as the reference level
- Create *L*-1 dummy variables for the remaining levels in the utility function

$$\text{Utility} = \ldots + \alpha_1 \cdot \text{ColourD}_1 + \alpha_2 \cdot \text{ColourD}_2 + \alpha_3 \cdot \text{ColourD}_3 + \ldots + \beta_1 \cdot \text{SideEffectsD}_1 + \beta_2 \cdot \text{SideEffectsD}_2 + \ldots$$

Colour	Coding		
	D_1	D_2	D_3
Red	1	0	0
Blue	0	1	0
Green	0	0	1
Yellow (ref)	0	0	0

Side effects	Coding	
	D_1	D_2
Mild (ref)	0	0
Moderate	1	0
Severe	0	1

Coding of categorical variables

Dummy coding

- Assume *L* levels (categories), select one level as the reference level
- Create *L*-1 dummy variables for the remaining levels in the utility function

$$\text{Utility} = \ldots + \alpha_1 \cdot \underset{0}{\text{ColourD}_1} + \alpha_2 \cdot \underset{1}{\text{ColourD}_2} + \alpha_3 \cdot \underset{0}{\text{ColourD}_3} + \ldots + \beta_1 \cdot \underset{0}{\text{SideEffectsD}_1} + \beta_2 \cdot \underset{0}{\text{SideEffectsD}_2} + \ldots$$

Colour	Coding		
	D_1	D_2	D_3
Red	1	0	0
Blue	0	1	0
Green	0	0	1
Yellow (ref)	0	0	0

Side effects	Coding	
	D_1	D_2
Mild (ref)	0	0
Moderate	1	0
Severe	0	1

Coding of categorical variables

Dummy coding

- Assume *L* levels (categories), select one level as the reference level
- Create *L*-1 dummy variables for the remaining levels in the utility function

$$\text{Utility} = \ldots + 0.4 \cdot \text{ColourD}_1 + 0.7 \cdot \text{ColourD}_2 - 0.3 \cdot \text{ColourD}_3 + \ldots - 0.2 \cdot \text{SideEffectsD}_1 - 0.7 \cdot \text{SideEffectsD}_2 + \ldots$$

Colour	Utility
Red	0.4
Blue	0.7
Green	-0.3
Yellow (ref)	0

Side effects	Utility
Mild (ref)	0
Moderate	-0.2
Severe	-0.7

Coding of categorical variables

Effects coding

- Assume *L* levels (categories), select one level as the reference level
- Create *L*-1 effects variables for the remaining levels in the utility function

$$\text{Utility} = \ldots + \alpha_1 \cdot \text{ColourE}_1 + \alpha_2 \cdot \text{ColourE}_2 + \alpha_3 \cdot \text{ColourE}_3 + \ldots + \beta_1 \cdot \text{SideEffectsE}_1 + \beta_2 \cdot \text{SideEffectsE}_2 + \ldots$$

Colour	Coding		
	E_1	E_2	E_3
Red	1	0	0
Blue	0	1	0
Green	0	0	1
Yellow (ref)	-1	-1	-1

Side effects	Coding	
	E_1	E_2
Mild (ref)	-1	-1
Moderate	1	0
Severe	0	1

Coding of categorical variables

Effects coding

- Assume *L* levels (categories), select one level as the reference level
- Create *L*-1 effects variables for the remaining levels in the utility function

$$\text{Utility} = \ldots + \alpha_1 \cdot \underset{0}{\text{ColourE}_1} + \alpha_2 \cdot \underset{1}{\text{ColourE}_2} + \alpha_3 \cdot \underset{0}{\text{ColourE}_3} + \ldots + \beta_1 \cdot \underset{-1}{\text{SideEffectsE}_1} + \beta_2 \cdot \underset{-1}{\text{SideEffectsE}_2} + \ldots$$

Colour	Coding		
	E_1	E_2	E_3
Red	1	0	0
→ Blue	0	1	0
Green	0	0	1
Yellow (ref)	-1	-1	-1

Side effects	Coding	
	E_1	E_2
→ Mild (ref)	-1	-1
Moderate	1	0
Severe	0	1

Coding of categorical variables

Effects coding

- Assume *L* levels (categories), select one level as the reference level
- Create *L*-1 effects variables for the remaining levels in the utility function

$$\text{Utility} = \ldots + 0.2 \cdot \text{ColourE}_1 + 0.5 \cdot \text{ColourE}_2 - 0.5 \cdot \text{ColourE}_3 + \ldots + 0.1 \cdot \text{SymptomsE}_1 - 0.4 \cdot \text{SymptomsE}_2 + \ldots$$

Colour	Utility
Red	0.2
Blue	0.5
Green	-0.5
Yellow (ref)	-0.2

Side effects	Utility
Mild (ref)	0.3
Moderate	0.1
Severe	-0.4

Coding of categorical variables

No need to change the database

- Can create dummy/effects coding using Boolean expressions when specifying utility functions in the model estimation script in Apollo

dummy coding

```
V = alpha_1 * (Colour=="Red")   +
    alpha_2 * (Colour=="Blue")  +
    alpha_3 * (Colour=="Green") +
    ...
    beta_1  * (SideEffects=="Moderate") +
    beta_2  * (SideEffects=="Severe")   +
    ...
```

effects coding

```
V = alpha_1 * ((Colour=="Red")   - (Colour=="Yellow")) +
    alpha_2 * ((Colour=="Blue")  - (Colour=="Yellow")) +
    alpha_3 * ((Colour=="Green") - (Colour=="Yellow")) +
    ...
    beta_1  * ((SideEffects=="Moderate") - (SideEffects=="Mild")) +
    beta_2  * ((SideEffects=="Severe")   - (SideEffects=="Mild")) +
    ...
```

Coding of categorical variables

Summary

- Categorical variables require numerical coding
- Dummy coding is most common and easiest, but other numerical coding schemes exist
- With *L* levels (categories), *L*-1 variables and parameters are required in the utility function

- Daly, A., T. Dekker, and S. Hess (2016) Dummy vs effects coding for categorical variables: Clarifications and extensions. *Journal of Choice Modelling*, 21, 36-41.