

VCD Package, by group Jeffreys

VCD in R

Visualizing Categorical Data

By Group: Jeffreys

Outline:

- structable
- Cotabplot
- Fourfold
- mosaic
- spine
- cd_plot

Introduction of VCD:

Categorical data: consists of variables whose values comprise a set of discrete categories.

The focus of this package is on visualization techniques and graphical methods designed to reveal patterns of relationships among categorical variables.

Utility functions: structable

Example:

```
structable(Dept+Gender ~ Admit, UCBA admissions)
```

Columns in nested
format

Dept	A		B		C		D		E		F	
Gender	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Admitted	512	89	353	17	120	202	138	131	53	94	22	24
Rejected	313	19	207	8	205	391	279	244	138	299	351	317

Counts in cells

The structable function can also return an object that you can pass to the mosaic function (discussed next)

Data: UCBAmissions

```
> structable(UCBAmissions)
```

Admit	Dept	Gender	Male	Female
Admitted	A		512	89
	B		353	17
	C		120	202
	D		138	131
	E		53	94
	F		22	24
Rejected	A		313	19
	B		207	8
	C		205	391
	D		279	244
	E		138	299
	F		351	317

Function: cotabplot

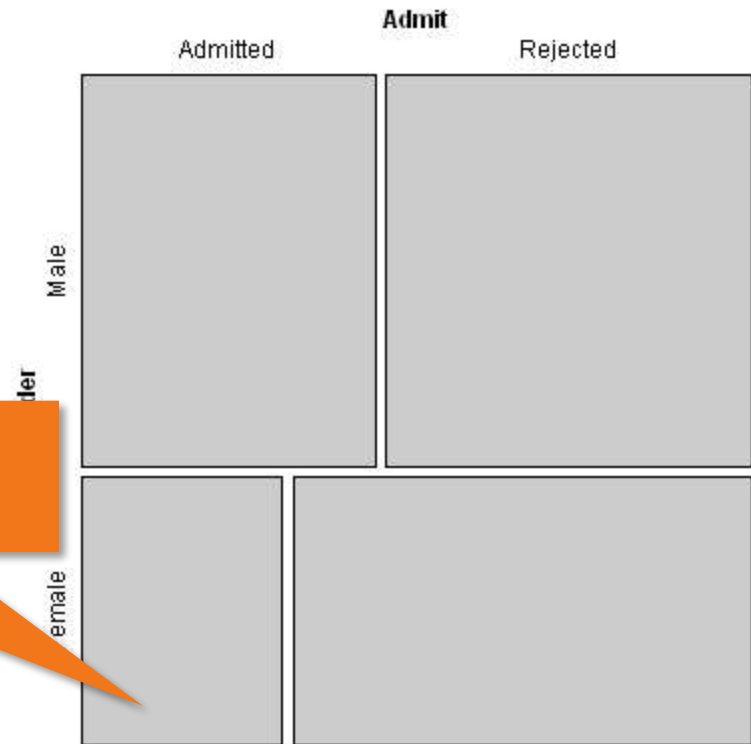
```
> GA<-margin.table(UCBAdmissions,c(2,1))
```

```
> structable(GA)
```

Gender	Admitted	Rejected
Male	1198	1493
Female	55	1278

```
> cotabplot(GA)
```

$P(\text{Admitted} \mid \text{Male}) > P(\text{Admitted} \mid \text{Female})$

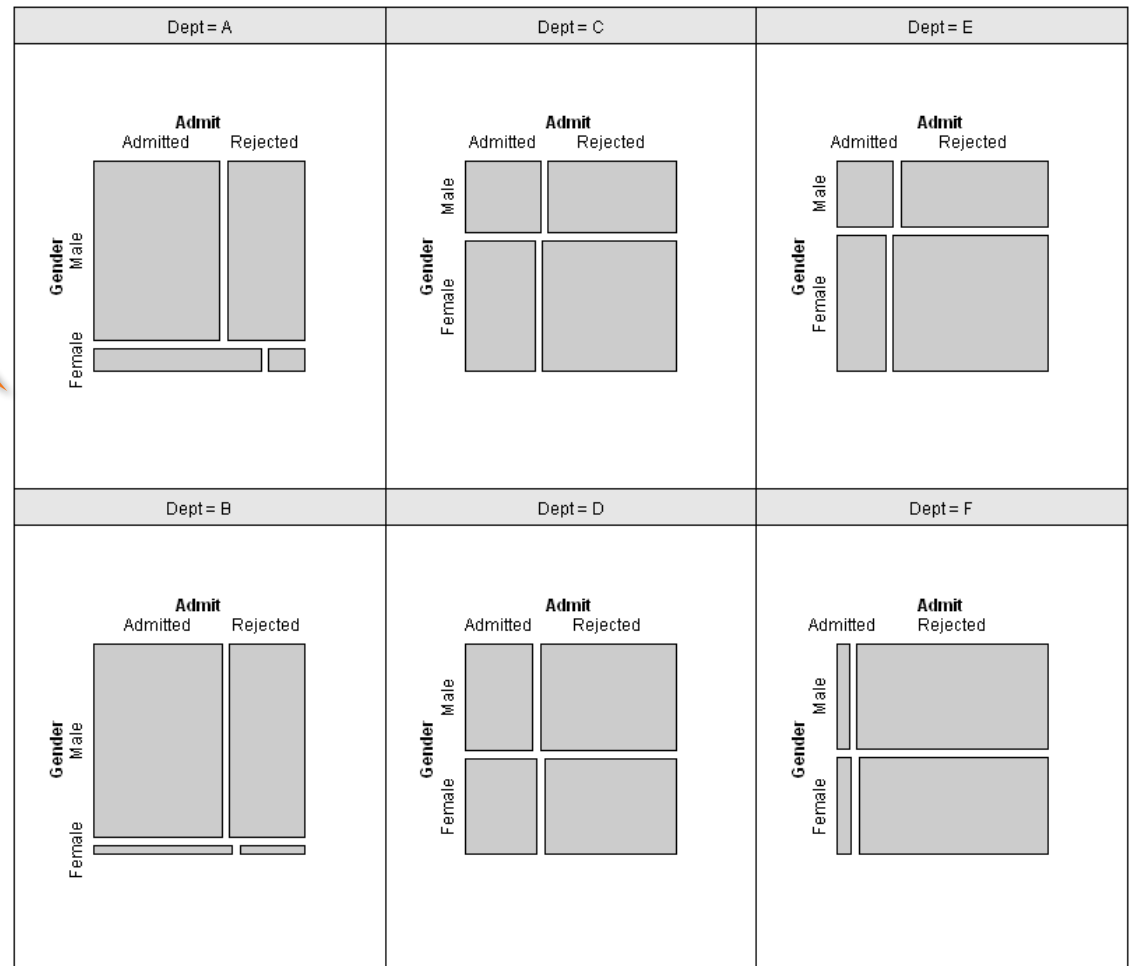


Function: cotabplot

```
> cotabplot(UCBAdmissions)
```

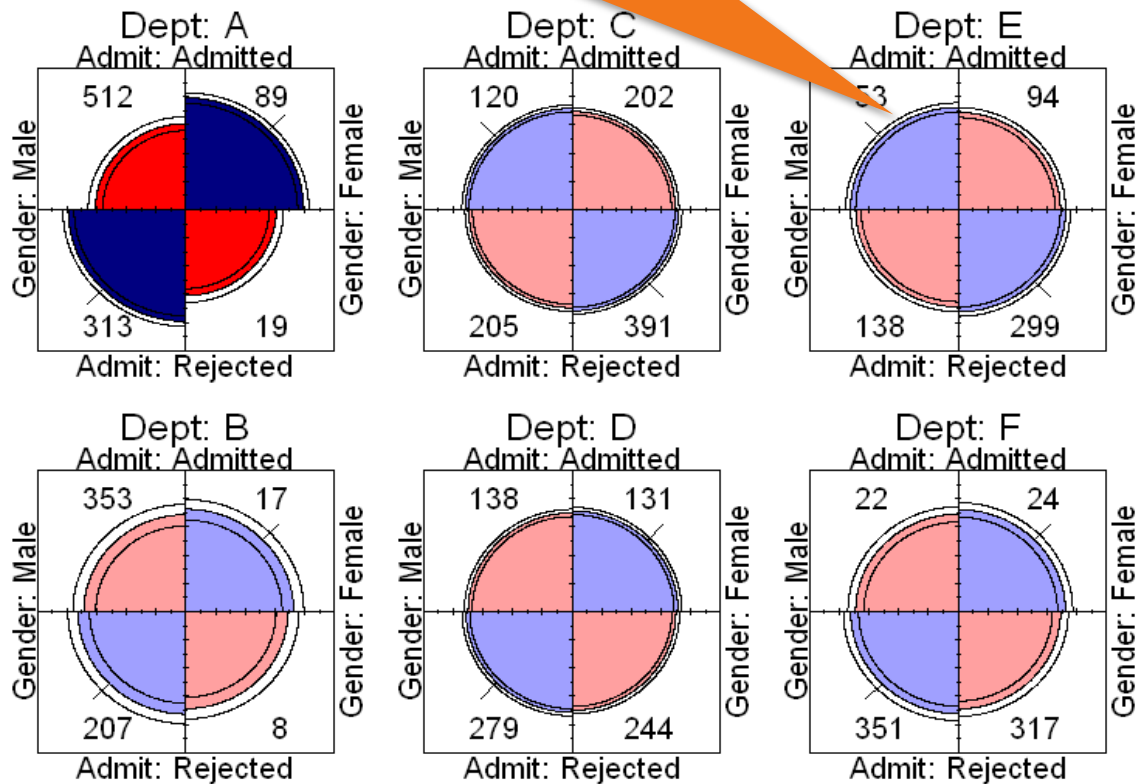
$P(\text{Admitted} \mid \text{Male}) ?$

$P(\text{Admitted} \mid \text{Female})$



Function: fourfold

radio $\sim \sqrt{n_{ij}}$, so its area \sim cell frequency



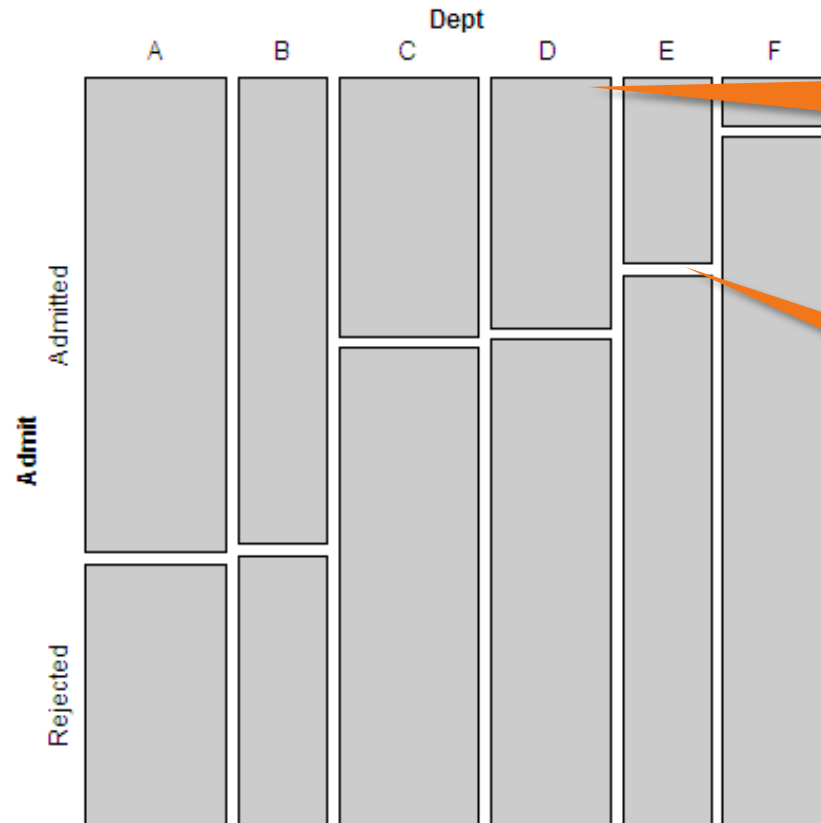
Display functions: mosaic

The 2-way mosaic table conveys the probability of each value of (a,b)

Example:

```
mosaic(~Dept + Admit, UCBA admissions, split_vertical=TRUE)
```

```
mosaic(structable(Admit ~ Dept, UCBA admissions), split_vertical=TRUE)
```



Marginal
probabilities across
the top

Conditional
probabilities within
each column

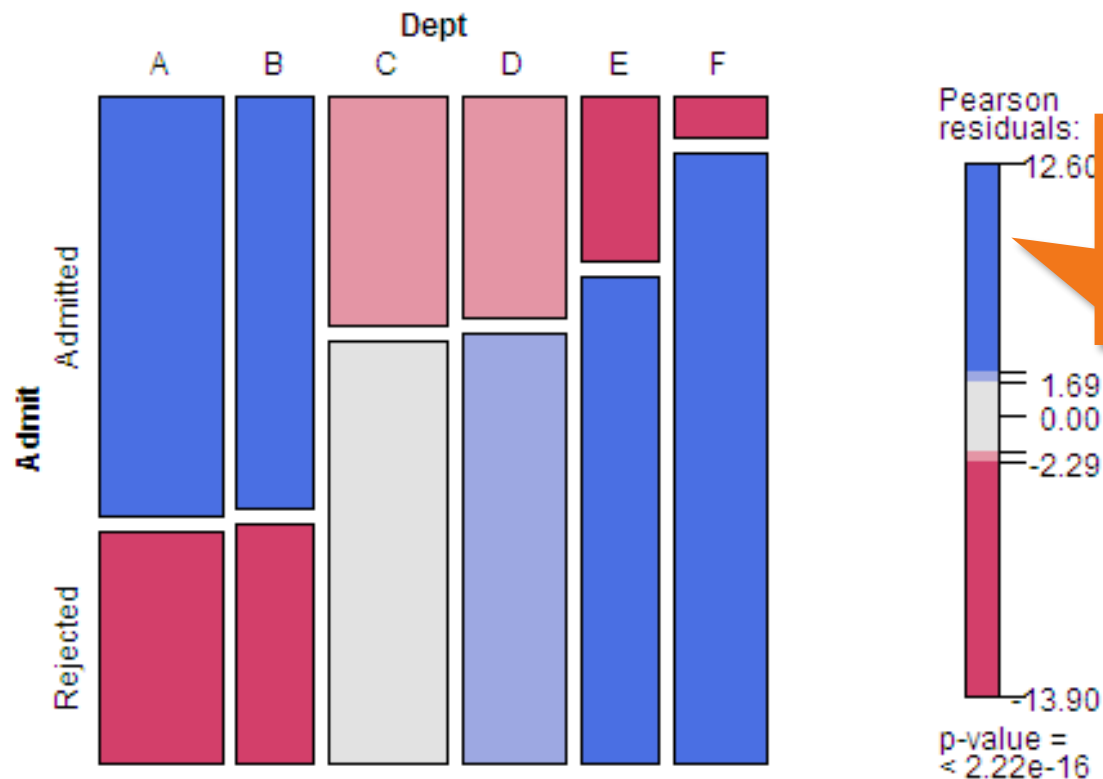
Display functions: mosaic

With shading we can see the residuals relative to a log-linear hypothesis

Example:

```
mosaic(~Dept + Admit, UCBAAdmissions, split_vertical=TRUE, gp=shading_max)
```

```
mosaic(structable(Admit ~ Dept, UCBAAdmissions), split_vertical=TRUE, gp=shading_max)
```



Residuals relative to hypothesis that $\text{prob} = \text{marginal } a1 \times \text{marginal } b1$

Display functions: mosaic

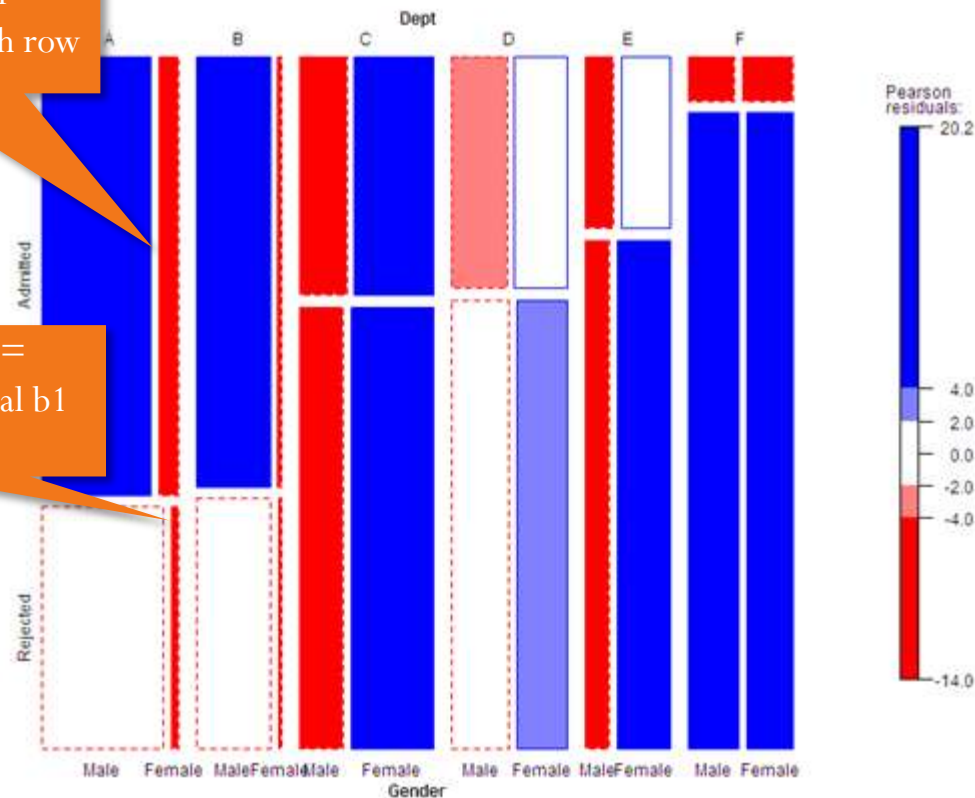
Multiway tables split the data into ever-finer cells

Example:

```
mosaic(~Dept+Admit+Gender, UCBA admissions, split_vertical=TRUE, gp=shading_Friendly)
```

Nested conditional probabilities within each row

Residuals for prob =
marginal a1 \times marginal b1
 \times marginal c1



Display functions: mosaic

Test for residuals against other hypotheses.

Example:
mosaic(A + B + C, data=..., split_vertical=TRUE, gp=shading_Friendly,
expected = ~A:B + C)

	Formula	Residual calculation based on:
H1: Complete independence	A + B + C	Prob (A, B, C) calculated based on: Prob {A B C}
H2: Joint independence	A:B + C	Prob (A, B, C) calculated based on: Prob {(A, B) C}
H3: Conditional independence	A:C + B:C	Prob (A, B C) calculated based on: Prob {(A, C) (B, C) A B}
H4: No 3-way interaction	A:C + B:C + A:B	Prob (A, B, C) calculated based on: Prob {(A, C) (B, C) (A, B) A B C}

These settings change the colouring, not the size and shape of the tiles themselves.

Data set interpret

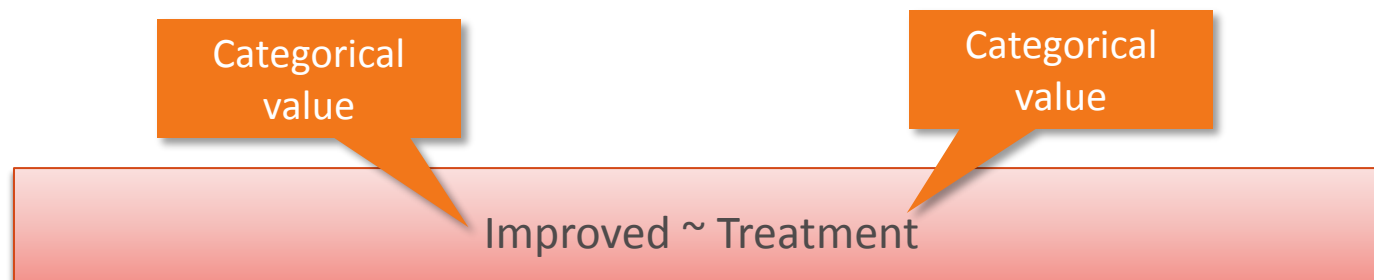
Arthritis: investigating a new treatment treatment for rheumatoid arthritis

- ID: patient ID.
- Treatment: factor indicating treatment (Placebo, Treated).
- Sex: factor indicating sex (Female, Male).
- Age: age of patient.
- Improved: ordered factor indicating treatment outcome (None, Some, Marked)

```
> head(Arthritis)
```

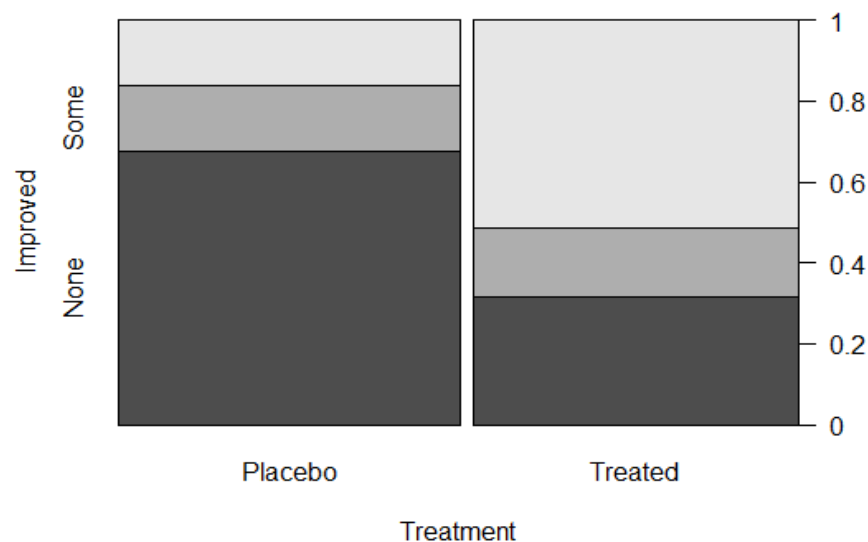
	ID	Treatment	Sex	Age	Improved
1	57	Treated	Male	27	Some
2	46	Treated	Male	29	None
3	77	Treated	Male	30	None
4	17	Treated	Male	32	Marked
5	36	Treated	Male	46	Marked
6	23	Treated	Male	58	Marked

Spine: Spine Plots and Spinograms



```
>(spine(Improved~Treatement, data = Arthritis))
```

Improved			
Treatment	None	Some	Marked
Placebo	29	7	7
Treated	13	7	21



Spine: Spine Plots and Spinograms

Categorical

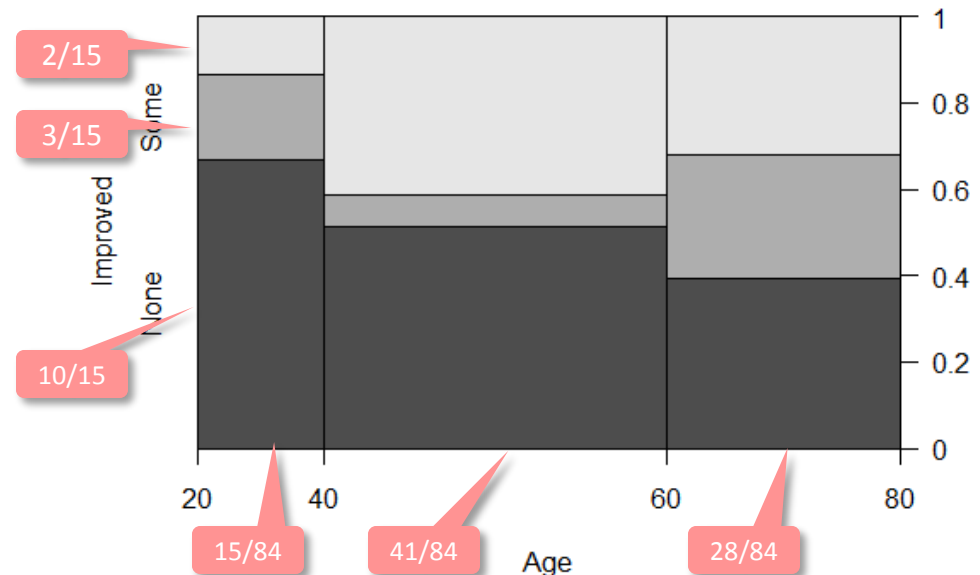
continuous

Improved ~ Age

- parameter passed to "hist"
- number, list, algorithm, fun
- Example: `quantile(Arthritis$Age)`

```
>(spine(Improved~Age, data = Arthritis, breaks=3))
```

Improved				
Age	None	Some	Marked	total
[20, 40]	10	3	2	15
(40, 60]	21	3	17	41
(60, 80]	11	8	9	28
total	42	14	28	84



➤ The widths of the bars corresponds to the relative frequencies of x.

➤ The heights of the bars correspond to the conditional relative frequencies of y in every x group

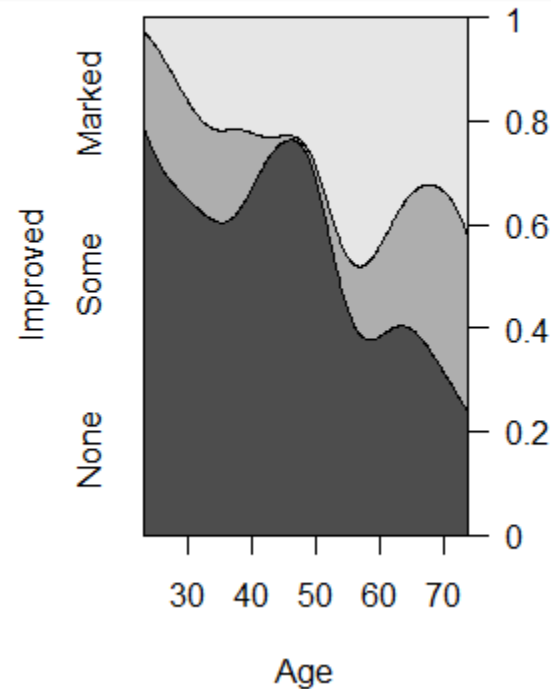
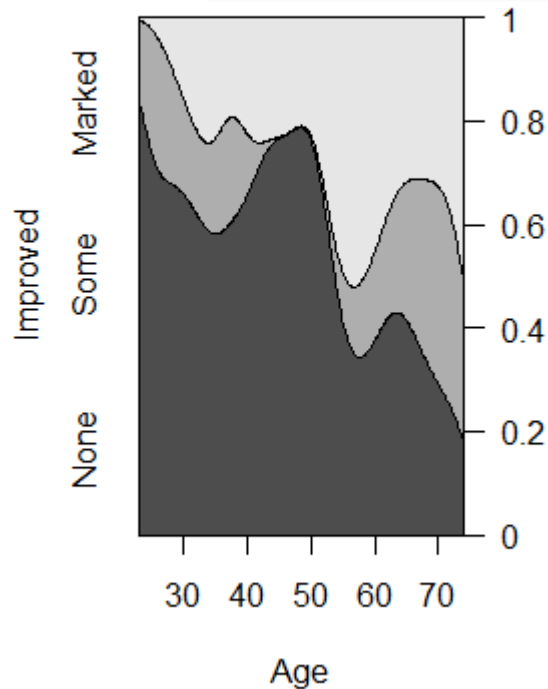
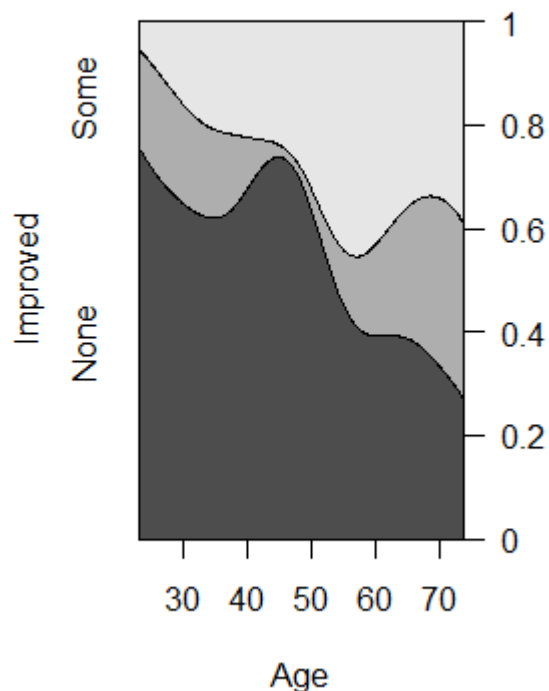
cd_plot: Conditional Density Plots

Description: Computes and plots conditional densities describing how the distribution of a categorical variable y changes over a numerical variable x .

```
> cd_plot(Improved ~ Age, data = Arthritis)
> cd_plot(Improved ~ Age, data = Arthritis, bw = 3)
> cd_plot(Improved ~ Age, data = Arthritis, bw = "SJ")
```

➤ bandwidth parameter passed to "density"
➤ Default: "nrd0", number, rule name

➤ Similar to spinograms, smoothing approach but not discrete
➤ The x axis is not a distorted x axis
➤ less reliable because of few observations



reference:

- kernel density estimate
http://en.wikipedia.org/wiki/Kernel_density_estimation
- The bandwidth of the kernel is a free parameter which exhibits a strong influence on the resulting estimate