Datenaufbereitung - Data Wrangling in R mit dplyr

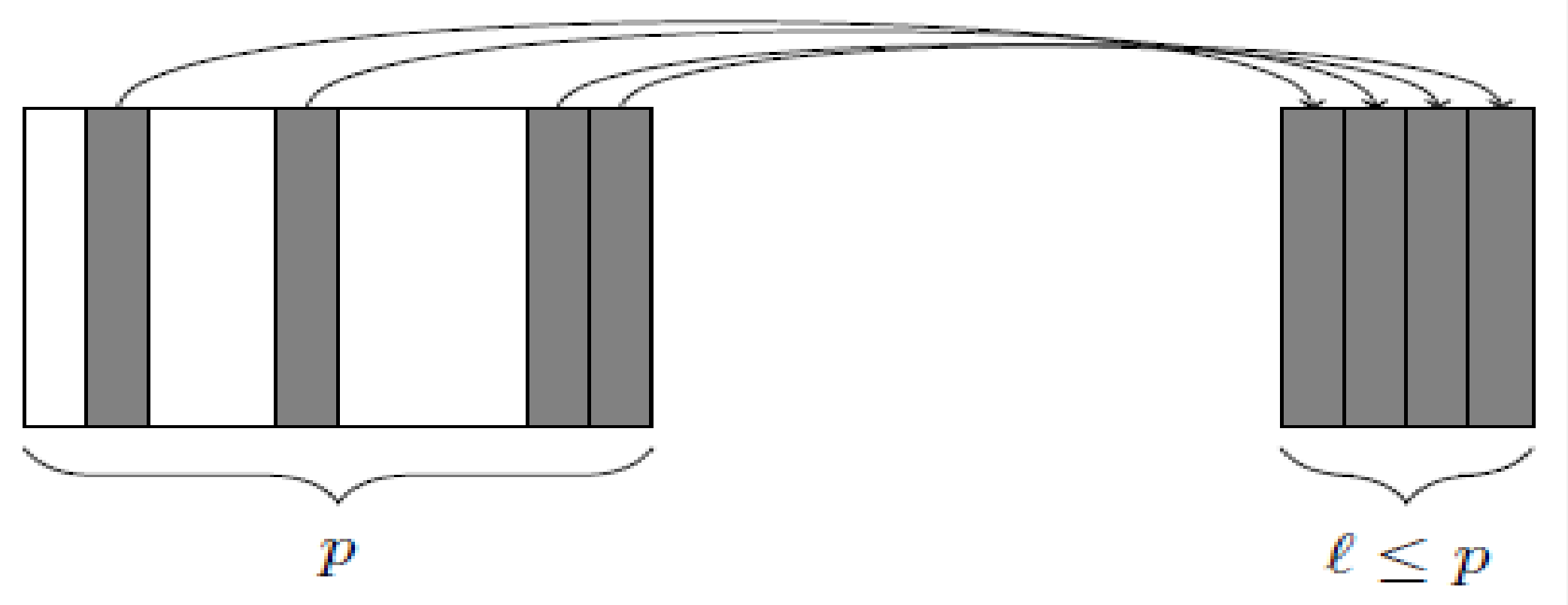
FH Technikum-Wien

Andreas Reschreiter

Datenaufbereitung – Data Wrangling – in R

- Data Wrangling ist die Aufbereitung von Rohdaten
- In R mit package `dplyr` mit den fünf wichtigen Befehlen:
  - `select()` für eine Teilmenge der Spalten (d. h. Features, Variablen)
  - `filter()` für eine Teilmenge der Zeilen (d. h. Beobachtungen)
  - `mutate()` fügt Spalten hinzu oder verändert vorhandene Spalten
  - `arrange()` sortiert die Zeilen (d. h. Beobachtungen)
  - `summarize()` aggregieren von Daten über Zeilen hinweg (z. B. Gruppieren)
- Alle fünf Befehle haben als **Input** und **Ouput** einen **data frame**
  - → beliebig kombinierbar (verschachtelbar)
  - → Mittels %>% Pipe sind Befehle sehr gut kombinierbar

Spalten auswählen select()



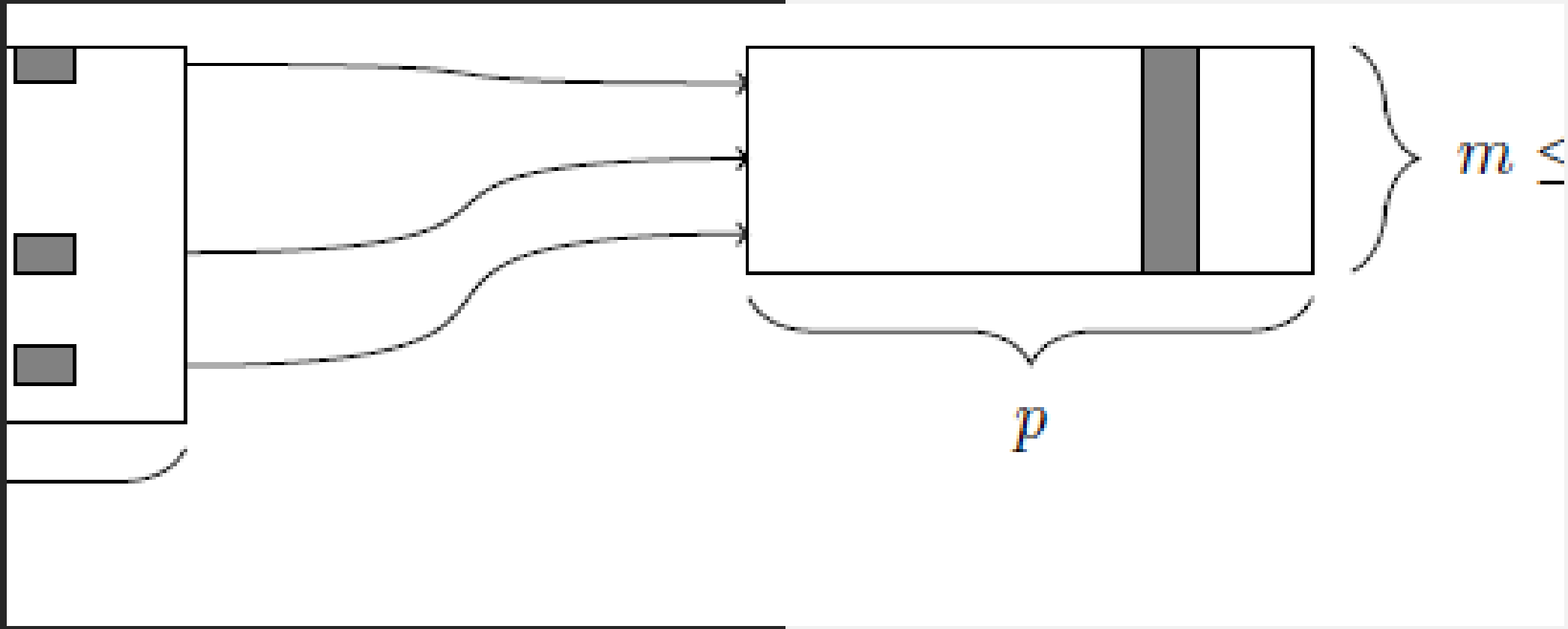
```
df
```

```
##    student semester note
## 1     Anna        1    1
## 2    Beate        3    2
## 3    Chris        5    1
## 4     Anna        1    4
## 5    Beate        3    5
## 6    Chris        5    1
## 7     Anna        1    3
## 8    Beate        3    3
## 9    Chris        5    3
```

```
select(df, student, semester)
```

```
##    student semester
## 1     Anna        1
## 2    Beate        3
## 3    Chris        5
## 4     Anna        1
## 5    Beate        3
## 6    Chris        5
## 7     Anna        1
## 8    Beate        3
## 9    Chris        5
```

Zeilen auswählen filter()



```
df
```

```
##    student semester note
## 1     Anna        1    1
## 2    Beate        3    2
## 3    Chris        5    1
## 4     Anna        1    4
## 5    Beate        3    5
## 6    Chris        5    1
## 7     Anna        1    3
## 8    Beate        3    3
## 9    Chris        5    3
```

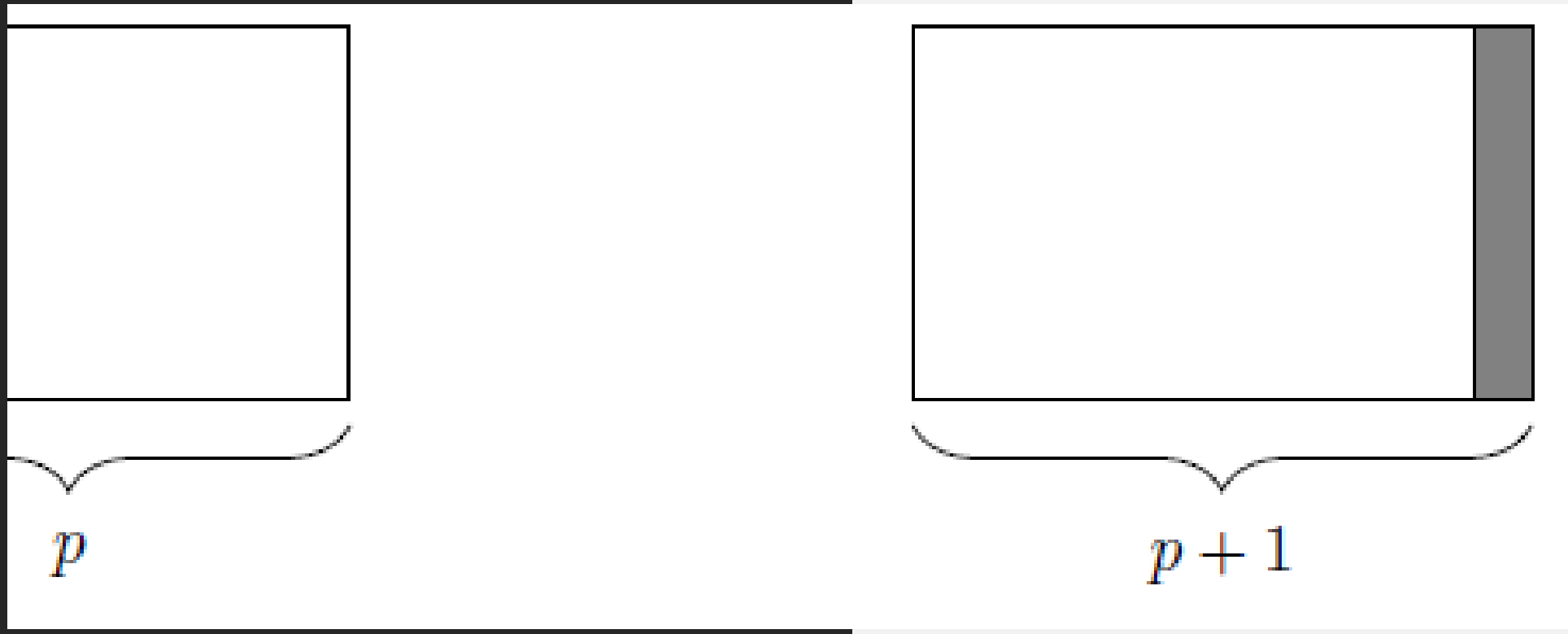
```
filter(df, note==1)
```

```
##    student semester note
## 1     Anna        1    1
## 2    Chris        5    1
## 3    Chris        5    1
```

```
filter(df, note==1 & student != "Anna")
```

```
##    student semester note
## 1    Chris        5    1
## 2    Chris        5    1
```

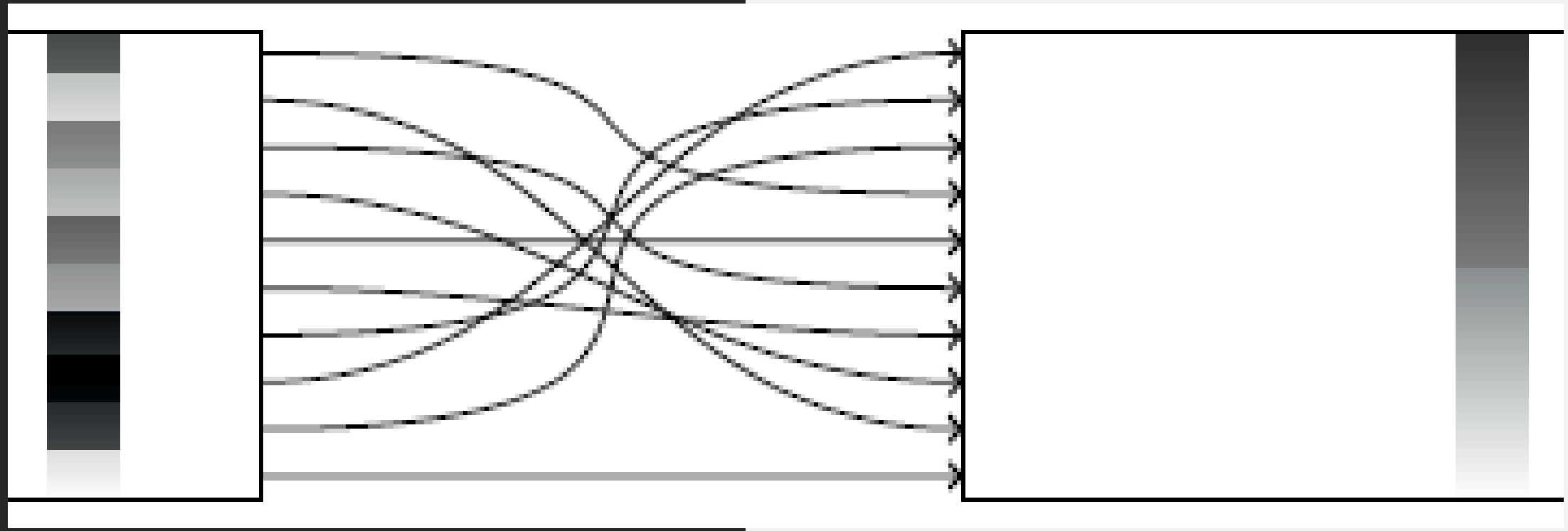
Spalten hinzufügen mutate()



```
df <-
mutate(df, textnote = ifelse(note==1, "sehr gut",
                             ifelse(note==2, "gut",
                                    ifelse(note==3, "befriedigend",
                                           ifelse(note==4, "genügend", "nicht genügend")
                                           )
                                    )
                             )
       )
df
```

```
##   student semester note         textnote
## 1    Anna        1    1         sehr gut
## 2   Beate        3    2              gut
## 3   Chris        5    1         sehr gut
## 4    Anna        1    4         genügend
## 5   Beate        3    5 nicht genügend
## 6   Chris        5    1         sehr gut
## 7    Anna        1    3     befriedigend
## 8   Beate        3    3     befriedigend
## 9   Chris        5    3     befriedigend
```

Zeilen sortieren arrange()



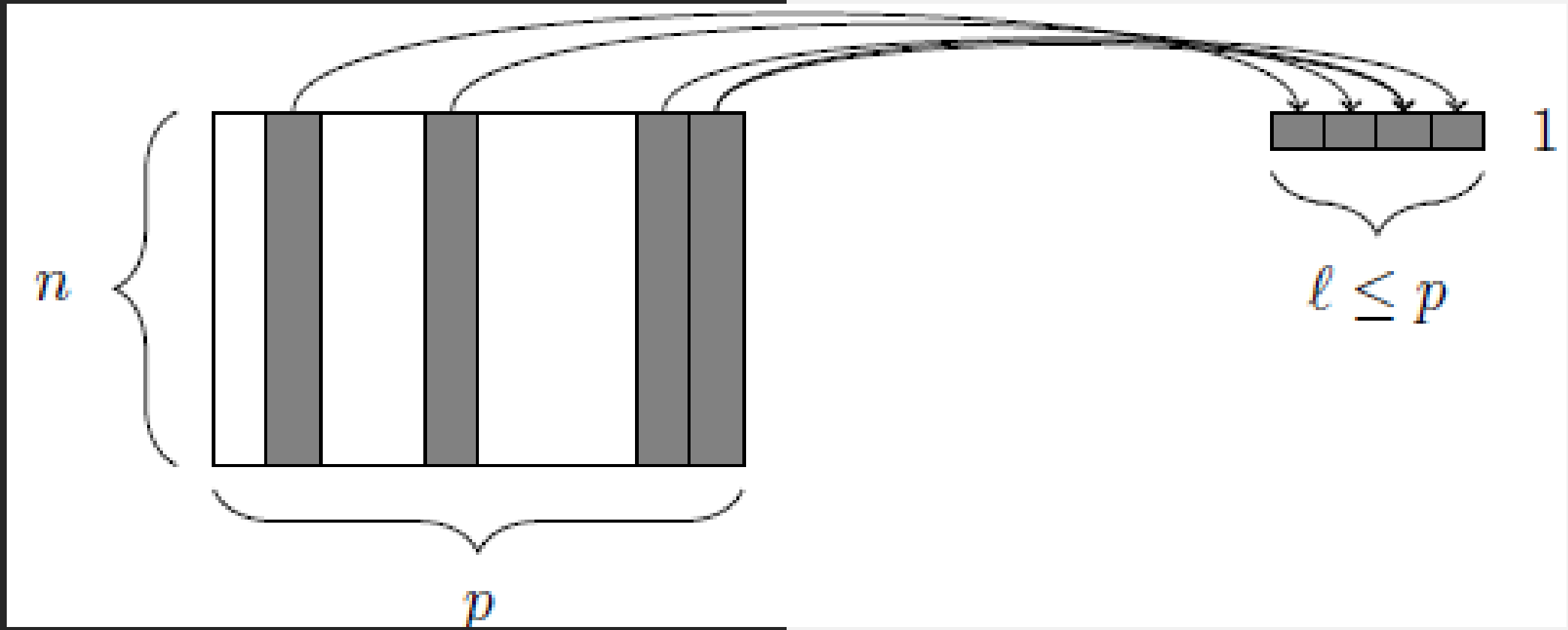
```
df
```

```
##   student semester note
## 1    Anna        1    1
## 2   Beate        3    2
## 3   Chris        5    1
## 4    Anna        1    4
## 5   Beate        3    5
## 6   Chris        5    1
## 7    Anna        1    3
## 8   Beate        3    3
## 9   Chris        5    3
```

```
arrange(df, note, desc(semester))
```

```
##   student semester note          textnote
## 1   Chris        5    1          sehr gut
## 2   Chris        5    1          sehr gut
## 3    Anna        1    1          sehr gut
## 4   Beate        3    2               gut
## 5   Chris        5    3    befriedigend
## 6   Beate        3    3    befriedigend
## 7    Anna        1    3    befriedigend
## 8    Anna        1    4         genügend
## 9   Beate        3    5 nicht genügend
```

Spaltenwert summarize()



```
df
```

```
##    student semester note
## 1     Anna        1    1
## 2    Beate        3    2
## 3    Chris        5    1
## 4     Anna        1    4
## 5    Beate        3    5
## 6    Chris        5    1
## 7     Anna        1    3
## 8    Beate        3    3
## 9    Chris        5    3
```

```
df %>% group_by(student) %>% summarise(
            N=n(),                          ## al
            notenschnitt= mean(note),
            best_note   = min(note),
            worst_note  = max(note)) %>%
  arrange(notenschnitt) %>%
  select(student, notenschnitt)
```

```
## # A tibble: 3 x 2
##    student notenschnitt
##    <fct>          <dbl>
## 1 Chris            1.67
## 2 Anna             2.67
## 3 Beate            3.33
```