HOCHSCHULE DER MEDIEN

Data Exploration in R

Basics

Prof. Dr. Jan Kirenz

HdM Stuttgart

Libraries

To use the code in this presentation, activate the following packages:

```r
library(tidyverse)
library(gt)
```

Data

- We use the dataset wage
- It contains wage and other data for a group of 3000 male workers.

```
library(tidyverse)

wage_df <- read_csv("https://raw.githubusercontent.com/kirenz/datasets/master/wage.csv")
```

- The data includes 3000 observations on 11 variables.

Data

- **year**: Year that wage information was recorded
- **age**: Age of worker
- **maritl**: A factor with levels:

  1. Never Married
  2. Married
  3. Widowed
  4. Divorced and
  5. Separated indicating marital status

- **race**: A factor with levels:

  1. White
  2. Black
  3. Asian and
  4. Other indicating race

- **education**: A factor with levels:

  1. < HS Grad
  2. HS Grad
  3. Some College
  4. College Grad and
  5. Advanced Degree indicating education level

- **jobclass**: A factor with levels:

  1. Industrial and
  2. Information indicating type of job

- **logwage**: Log of workers wage
- **wage**: Workers raw wage

Counts and Tables

You should use this method if the data is: Categorical

In this section you will learn how to do data exploration for categorical variables using tables (also called contingeny tables) and counts.

Count

- Use data wage_df.
- Perform count() on maritl
- Sort the values.
- Use gt() to print the table.

```
      %>%
    (     ,
        = TRUE) %>%
    ()
```

maritl	n
2. Married	2074
1. Never Married	648
4. Divorced	204
5. Separated	55
3. Widowed	19

Count

- Use data wage_df.
- Perform count() on maritl and education
- Sort the values.
- Use gt() to print the table.

Count the combined appearences of maritl and education and sort the values:

```
    %>%
  (   ,     ,
       = TRUE) %>%
  ()
```

maritl	education	n
2. Married	2. HS Grad	651
2. Married	4. College Grad	487
2. Married	3. Some College	421
2. Married	5. Advanced Degree	341
1. Never Married	2. HS Grad	219
2. Married	1. < HS Grad	174

Count

- Use data wage_df.
- Obtain the sum of the quantitative variable wage for the different levels of maritl
- Name the new variable "Sum".
- Use gt() to print the table.

Read the documentation for count() to learn how to perform this task.

```
    %>%
   (   ,
      = wage,
        = "Sum") %>%
  ()
```

maritl	Sum
1. Never Married	60092.052
2. Married	246516.180
3. Widowed	1891.234
4. Divorced	21044.489
5. Separated	5566.868

Total counts

- Total counts are an useful way to represent the observations that fall into each combination of the levels of categorical variables.

- We create a contingency table of the two categorical variables jobclass and race and call the result tab:

```
tab <- table(wage_df$jobclass, wage_df$race)

kable(tab)
```

	1. White	2. Black	3. Asian	4. Other
1. Industrial	1325	111	86	22
2. Information	1155	182	104	15

Joint proportions

- We can also view the percentage of each cell in relation to the total amount of all observations (here n = 3000).
- Therefore, you have to simply divide the numbers from our total counts with 3.000.

The following code generates tables of *joint* proportions:

```
prop <- prop.table(tab)*100

kable(prop, digits = 2)
```

	1. White	2. Black	3. Asian	4. Other
1. Industrial	44	3.7	2.9	0.73
2. Information	38	6.1	3.5	0.50

For example, around ?? % of all people in the dataset are white industrial workers.

Conditional proportions

- You also may want to know the probability that workers have a certain jobclass, given that they have a particular ethnical background.
- This is a so called **conditional probability**.
- Conditional probability represents the chance that one event will occur given that a second event has already occurred.

Conditional proportions

- The following code generates tables of *conditional* proportions:

```
# conditional on columns
prop_col <- prop.table(tab, 2)*100

kable(prop_col, digits = 2)
```

	1. White	2. Black	3. Asian	4. Other
1. Industrial	53	38	45	59
2. Information	47	62	55	41

We performed a columnwise evaluation and are now able to answer the following question:

- Approximately what proportion of all white workers are industrial workers? The answer is: around ?? %.

Conditional proportions: rows

Now we want to obtain the probability that workers have a certain race, given their jobclass.

```r
# conditional on rows
prop.table(tab, 1)
```

```
##                
##                    1. White   2. Black   3. Asian   4. Other
##   1. Industrial  0.85816062 0.07189119 0.05569948 0.01424870
##   2. Information 0.79326923 0.12500000 0.07142857 0.01030220
```

We performed a rowwise evaluation and are now able to answer the following question:

- Approximately what proportion of all industrial workers are white?
- The answer is: around 86%.

Chi-squared Test of Independence

Finally, let's test the hypothesis whether the variable jobclass is independent of the variable race at .05 significance level.

```
chisq.test(tab)
```

```
## 
##     Pearson's Chi-squared test
## 
## data:  tab
## X-squared = 29.331, df = 3, p-value = 1.908e-06
```

As the p-value is smaller than the .05 significance level, we reject the null hypothesis that the jobclass is independent of the race of the workers.

HOCHSCHULE
DER MEDIEN

Thanks!

Prof. Dr. Jan Kirenz

HdM Stuttgart
Nobelstraße 10
70569 Stuttgart