Elements of Statistics

I Chapter 2: Methods of Data Collection and Visualisation

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier
Faculty IV
Economic and Social Statistics Department

Winter term 2021/22

Notes

Eurostat database

Indicators for:

- Key indicators on EU policy
- General and regional statistics
- Economics and finance
- Population and social conditions
- Industry, trade and services
- Agriculture, forestry and fisheries
- International trade
- Transportation
- Environment and Energy
- Science, technology, digital technology

https://ec.europa.eu/eurostat/data/database

Notes

European Social Survey (ESS)

- Introduced in 2001
- International **Survey** in Europe (more than 30 countries)
- *Academically driven*
- Surveyed every 2 years
- Measures the attitudes, beliefs and behaviour patterns
- Chart stability and change in social structure
- Introduce soundly-based indicators

Homepage:
https://www.europeansocialsurvey.org/

Documents and data files for German survey:
https://www.europeansocialsurvey.org/data/country.html?c=germany

Notes

Eurosystem Household Finance and Consumption Survey (HFCS)

- Initiated in 2006
- Currently 19 countries of the Eurozone plus Croatia, Hungary and Poland
- Initiative of the European Central Bank (ECB)
- Approx. surveyed every 3 years
- Data on households:
  - Real assets and their financing
  - Liabilities
  - Income
  - Consumption
  - Socio economic information
  - socio demographic information
- High relevance for monetary and fiscal policy (see financial crisis)

http://www.ecb.int/home/html/researcher_hfcn.en.html

Notes

Terminology

Wirtschafts- und Sozialstatistik Universität Trier

POP	Population
n	Number of units to be analysed; enumerated by $i = 1, \ldots, n$ (later on: N for POP and n for sample)
m	Number of categories or ranks
x_i	Value of variable X for i-th unit ($i = 1, \ldots, n$)
n_j	Absolute frequency of j-th category or j-th rank ($j = 1, \ldots, m$)
p_j	Relative frequency of j-th category or j-th rank ($j = 1, \ldots, m$)

Notes

Frequency distribution

Wirtschafts- und Sozialstatistik Universität Trier

The frequency distribution of a variable summarises its categories or ranks and the related frequencies. It can be determined in an absolute or relative sense and is presented in a frequency table.

We have

$$\sum_{j=1}^{m} n_j = n \qquad \text{and} \qquad \sum_{j=1}^{m} p_j = \sum_{j=1}^{m} \frac{n_j}{n} = 1 \, .$$

Notes

Example 2.1: Unemployment (1)

Wirtschafts- und Sozialstatistik Universität Trier

In an attempt to estimate the number of unemployed people in Germany, we first analyse the one-dimensional (univariate) frequency distributions.

- Distribution by gender:

male	female	$\sum$
666	524	1,190

Realisation in R:

```
setwd("path") # Choose your working directory on your own.
load("Example2-1.RData")
table(Unemployment$Gender)
```

```
Male  Female
  666    524
```

```
length(Unemployment$Gender)
```

```
[1] 1190
```


Notes

Example 2.2: Fields of study (1)

Wirtschafts- und Sozialstatistik Universität Trier

Students in the course *Mathematical Statistics* were asked about their field of study. The following original list emerged:

ECO, ECO, ECO, BA, SOC, MAT, BMAT, BMAT, ECO, ECO, ECO, SOC, SOC, ECO, ECO, SOC, MAT, CS, ECO, ECO.

Field of study	j	Tally	n_j	p_j (in %)
Business administration	1	|	1	5
Economics	2	~~||||~~ ~~||||~~	10	50
Sociology	3	||||	4	20
Mathematics	4	||	2	10
Business mathematics	5	||	2	10
Computer sciences	6	|	1	5
			20	100

Notes

Example 2.2: Fields of study (2)

Determination of frequency tables in R:

```
x2_2 <- factor(c("ECO", "ECO", "ECO", "BA", "SOC", "MAT",
                 "BMAT", "BMAT", "ECO", "ECO", "ECO", "SOC",
                 "SOC", "ECO", "ECO", "SOC", "MAT", "CS",
                 "ECO", "ECO"),
        levels = c("BA", "ECO", "SOC", "MAT", "BMAT", "CS")
)
n_j <- table(x2_2)
p_j <- prop.table(n_j)
n_j
x2_2
 BA ECO  SOC MAT BMAT CS
  1  10   4   2   2   1
p_j
x2_2
 BA   ECO  SOC  MAT  BMAT  CS
0.05 0.50 0.20 0.10 0.10 0.05

p_j * 100
x2_2
BWL VWL SOZ MAT WIM INF
  5  50  20  10  10   5
```


Notes

Grouping of metric variables

- Variables with few unique values can be treated as before (e.g. number of children in household)
- Variables with many unique values need grouping (e.g. income)

→ Splitting of range $[x_0^o; x_m^o]$ into m classes:

x_j^o	Upper boundary of j-th class
x_0^o	Lower boundary of first class
$x_j' = \frac{1}{2}(x_j^o + x_{j-1}^o)$	Mid of j-th class
$n_j(p_j)$	Number (share) of observations in j-th class

The i-th observation x_i falls into the j-th class if $x_{j-1}^o \leq x_i < x_j^o$.

Notes

Example 2.3: compare Example 2.1 (1)

- Distribution of unemployed by age class:

[0; 15)	[15; 25)	[25; 45)	[45; 65)	≥ 65	$\sum$
0	190	523	475	2	1.190

Illustration of absolute frequency of classes $m = 2$ (Ex. 2.1) resp. $m = 5$ (here).

Application in R:

```
x_o       <- c(0, 15, 25, 45, 65, Inf)
age_class <- cut(x = Unemployment$Age, breaks = x_o,
                 right = FALSE)
table(age_class)
length(age_class)

age_class
  [0,15)  [15,25)  [25,45)  [45,65) [65,Inf)
       0      190      523      475        2

length(age_class)

[1] 1190
```


Notes

Example 2.3: compare Example 2.1 (2)

Further analysis of the unemployment dataset in R:

```
attach(Unemployment)
summary(Income)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  247.0   564.9   644.4   642.9   723.1   946.3
```

- What would be the result for other surveys?
- Which influence do *cut* incomes have?
- How exact are the observations?

```
x_o_new <- c(200, 350, 500, 650, 800, Inf)
income_class <- cut(x = Income, breaks = x_o_new,
                    right = FALSE
)
summary(income_class)
[200,350) [350,500) [500,650) [650,800) [800,Inf)
        5       140       480       451       114
```


Notes

Some visualisation tools

- Horizontal bar chart
  - Absolute
  - Relative
- Vertical bar chart
  - Absolute
  - Relative
- Pie chart
- Spider plot
- Histogram
- Sum function

You should always consider the scaling of the data about to be visualised!

Notes

Example 2.4: German construction activity (1)

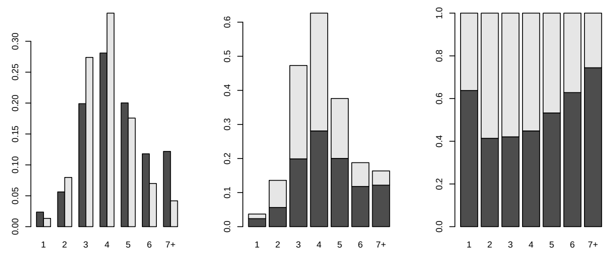
Object	Units	West	East	Total
West: former federal territory		East: new federal states and Berlin		
Housing stock 2011 (available figures)				
Flats	1 000	31 585.2	8 888.7	40 473.8
with ... rooms				
1	1 000	738.7	120.0	858.7
2	1 000	1 784.9	698.5	2 483.5
3	1 000	6 222.7	2 405.7	8 628.4
4	1 000	8 765.1	3 043.7	11 808.8
5	1 000	6 341.9	1 585.1	7 927.1
6	1 000	3 797.6	644.3	4 442.0
7 and more	1 000	3 934.2	391.3	4 325.5
Rooms in total	1 000	143 321.5	35 686.2	179 007.6
Difference to 2004	1 000	718.0	71.3	789.2
Living space in total	million sqm	2 862.1	654.1	3 516.2

Source (15/10/2012): See Bautätigkeit, Wohnungsbestand on
`https://www.destatis.de/DE/ZahlenFakten/Wirtschaftsbereiche/Bauen/Bautaetigkeit/Bautaetigkeit.html`

Notes

Example 2.4: German construction activity (2)

Graphical presentation with bar charts (see table:



```
round(App_pj, digits = 4) # See next slide!
```

```
          1      2      3      4      5      6     7+
West 0.0234 0.0565 0.1970 0.2775 0.2008 0.1202 0.1246
East  0.0135 0.0786 0.2707 0.3424 0.1783 0.0725 0.0440
```


Notes

Example 2.4: German construction activity (3)

Bar chart of the previous slide in R:

```
load("Example2-4.RData")

App_pj <- t(apply(Housing[, 2:3], MARGIN = 2,
                  FUN = prop.table))
colnames(App_pj) <- Housing$Number_of_rooms

par(mfrow = c(1, 3))
barplot(App_pj, beside = TRUE)
barplot(App_pj)
barplot(apply(App_pj, 2, prop.table))
par(mfrow = c(1, 1))
```

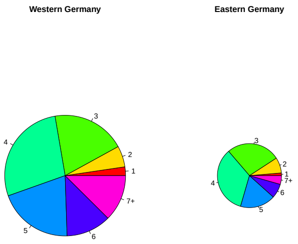
You can create vertical bar charts by setting the argument `horiz` of the `barplot` function to `TRUE`.

Notes

Example 2.4: German construction activity (4)

Wirtschafts- und Sozialstatistik
Universität Trier

The intention here is to illustrate the number of flats subject to the number of rooms per flat and compare the relevant figures for Western and Eastern Germany (see table above).



Notes

Example 2.4: German construction activity (5)

Wirtschafts- und Sozialstatistik
Universität Trier

Pie charts of the previous slide in R:

```
sum_west <- sum(Housing$West)
sum_east <- sum(Housing$East)
radius_west <- min(1, sqrt(sum_west / sum_east))
radius_east <- min(1, sqrt(sum_east / sum_west))

par(mfrow = c(1, 2))
pie(Housing$West,
    col = rainbow(n = 7),
    radius = radius_west,
    labels = Housing$Number_of_rooms
    )
title("Western Germany", line = -4)
pie(Housing$East,
    col = rainbow(n = 7),
    radius = radius_east,
    labels = Housing$Number_of_rooms
    )
title("Eastern Germany", line = -4)
par(mfrow = c(1, 1))
```


Notes

Sustainable Development Goals (SDG)

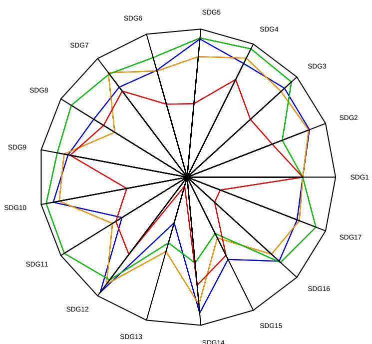
Wirtschafts- und Sozialstatistik
Universität Trier

Siehe Sustainable Development Goals Knowledge Platform:
`https://sustainabledevelopment.un.org/?menu=1300`

Notes

Spider plot

Wirtschafts- und Sozialstatistik
Universität Trier



Legend: GER, FRA, NOR, RUS

Notes

Grouping of metric variables

Wirtschafts- und Sozialstatistik Universität Trier

- Number of classes
  - 5 – 20
  - Rule of thumb: $\sqrt{n}$
    (problematic for official statistics; e.g. micro census or census)
- Location of classes
  - A higher denseness of observations should lead to narrower class intervals
  - Only few differing class widths
  - Class widths and class mids should rather be integers
  - No open marginal classes

Some special cases might have to be accounted for explicitly.

Notes

The histogram

Wirtschafts- und Sozialstatistik Universität Trier

Starting from a given grouping of a metric variable without any open classes, the areas of the rectangles drawn above the class intervals $[x^o_{j-1}; x^o_j)$ are matched to the relative frequencies p_j.

The rectangles' heights h_j are calculated by rearranging $p_j = d_j \cdot h_j$, with $d_j = x^o_j - x^o_{j-1}$ as the class width.

Notes

The empirical distribution function

Wirtschafts- und Sozialstatistik Universität Trier

The empirical distribution function specifies the share of observations which exhibit a value less than or equal to x.
It holds that:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(x_i \leq x) \quad , \qquad \text{with} \qquad \mathcal{I}(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{else} \end{cases}$$

as the indicator function.

Notes

Some remarks (1)

Wirtschafts- und Sozialstatistik Universität Trier

- The subscript n, giving the size of the population/sample, may be dropped due to redundance.
- Instead of the empirical distribution function (also called the cumulative relative frequency distribution), the cumulative absolute frequency distribution is used occasionally.
  Then, the following holds: $F^*_n(x) = n \cdot F_n(x)$.
- The empirical distribution function requires at least an ordinal variable. An interpretation of distances on the abscissa is only possible for metric variables.

Notes

Some remarks (2)

- When we start with grouped data, information on concrete values within classes is typically missing. Nevertheless, by using the class boundaries we may still calculate the cumulative relative frequencies. Given the *assumption* that the values are uniformly distributed within the classes, we can connect the cumulative relative frequencies at the class boundaries $F(x_j^o)$ as a polygonal line, thereby connecting the following points: $(x_0^o, 0); (x_1^o, F(x_1^o)); \ldots; (x_{m-1}^o, F(x_{m-1}^o)); (x_m^o, 1)$.

  Then, within the j-th class $(x_{j-1}^o \leq x < x_j^o)$ we have

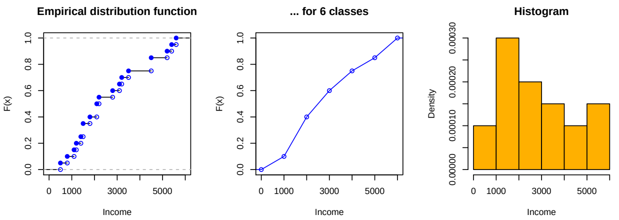
$$F(x) = F(x_{j-1}^o) + p_j \cdot \frac{x - x_{j-1}^o}{x_j^o - x_{j-1}^o} \quad .$$

Notes

Example 2.5: Income (1)

In an income study with a sample size of $n = 20$, the following values were recorded: 3500, 3200, 2100, 500, 1800, 2100, 5600, 4500, 1400, 1200, 1500, 2200, 3100, 1500, 2800, 1100, 5200, 4500, 5400, 800.

The resulting empirical distribution function (original and grouped data) as well as the histogram with class boundaries of 0, 1000, 2000, 3000, 4000, 5000 and 6000 are shown here:



Notes

Example 2.5: Income (2)

Graphics of the previous slide in R:

```
x2_5 <- c(3500,3200,2100,500,1800,2100,5600,4500,1400,1200,
          1500,2200,3100,1500,2800,1100,5200,4500,5400,800)
ant <- (1:length(x2_5)) / length(x2_5)

par(mfrow=c(1,3))
plot(ecdf(x2_5), xlab = "Income",
     ylab = expression(F[n](x)),
     main = "Empirical distribution function",
     col = "blue")
points(sort(x2_5), c(0, ant[-length(ant)]), col = "blue")

x_o <- c(0,1000,2000,3000,4000,5000,6000) # Overwriting
x2_5_kl <- cut(x2_5, x_o, right = FALSE)
F_j <- cumsum(prop.table(table(x2_5_kl)))
plot(x = c(0,0), main = "...for 6 classes",
     xlab = "Income", ylab = expression(F[n](x)),
     xlim = c(0,6000), ylim = c(0,1), type = "n")
lines(x = x_o, y = c(0,F_j), col = "blue")
points(x = x_o, y = c(0,F_j), col = "blue")
```


Notes

Example 2.5: Income (3)

```
hist(x2_5, probability = TRUE, xlab = "Income",
     ylab = "Density", main = "Histogram", col = "#FFB000")
par(mfrow = c(1, 1))
```


Notes

Wirtschafts- und Sozialstatistik Universität Trier

Example 2.6: Time to completion (1)

For $n = 200$ pupils, the time taken to solve a problem was recorded:

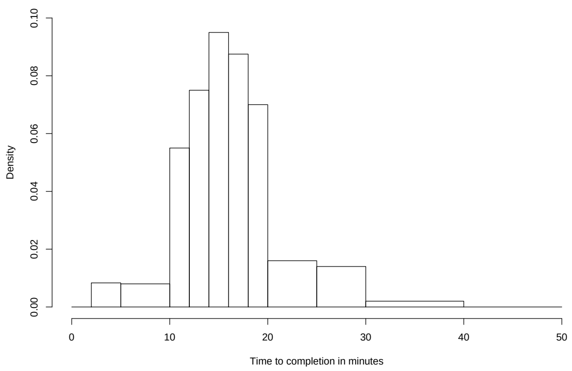
Class (min.)	j	n_j	p_j	d_j	h_j	$\sum_{\nu=1}^{j} p_\nu$
2 up to, but less than 5	1	5	0.025	3	0.0083	0.025
5 up to, but less than 10	2	8	0.040	5	0.0080	0.065
10 up to, but less than 12	3	22	0.110	2	0.0550	0.175
12 up to, but less than 14	4	30	0.150	2	0.0750	0.325
14 up to, but less than 16	5	38	0.190	2	0.0950	0.515
16 up to, but less than 18	6	35	0.175	2	0.0875	0.690
18 up to, but less than 20	7	28	0.140	2	0.0700	0.830
20 up to, but less than 25	8	16	0.080	5	0.0160	0.910
25 up to, but less than 30	9	14	0.070	5	0.0140	0.980
30 up to, but less than 40	10	4	0.020	10	0.0020	1.000
$\sum$		200	1.000			

See **Schaich, E.: Schätz- und Testmethoden für Sozialwissenschaftler (1998), 3rd edition, Vahlen, p. 17 ff.**

Notes

Wirtschafts- und Sozialstatistik Universität Trier

Example 2.6: Time to completion (2)



Notes

Wirtschafts- und Sozialstatistik Universität Trier

Example 2.6: Time to completion (3)

Histogram of the previous slide in R:

```
load("Example2-6.RData")

x_o <- c(2,5,10,12,14,16,18,20,25,30,40)
hist(Time, breaks = x_o,
     right = FALSE, main = "",
     xlim = c(0,50), ylim = c(0,0.1),
     xlab = "Time to completion in minutes", ylab = "Density")
```

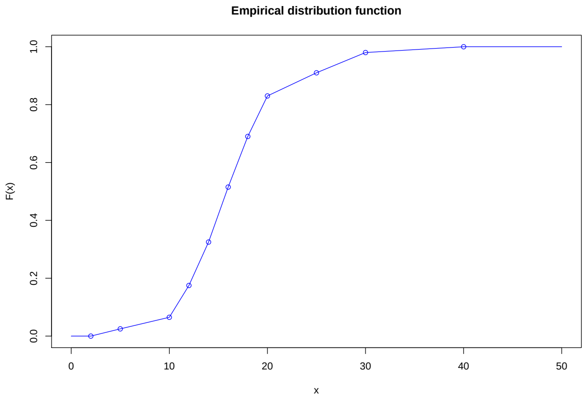
Empirical distribution function of the next slide in R:

```
F_j <- cumsum(prop.table(table(cut(Time, x_o,
                                   right = FALSE))))
plot(x = c(0,0), type = "n",
     main = "Empirical distribution function",
     ylab = "F(x)", xlab = "x", xlim = c(0,50),
     ylim = c(0,1))
lines(x = c(0,x_o,50), y = c(0,0,F_j,1), col = "blue")
points(x = c(0,x_o,50), y = c(0,0,F_j,1), col = "blue")
```


Notes

Wirtschafts- und Sozialstatistik Universität Trier

Example 2.6: Time to completion (4)



Notes

Kernel density estimation

Instead of histograms, which show discontinuities at class boundaries, *approximations* may be used.

Kernel density estimator

For a given kernel $K(u)$ and the data $x_1, \ldots, x_n$

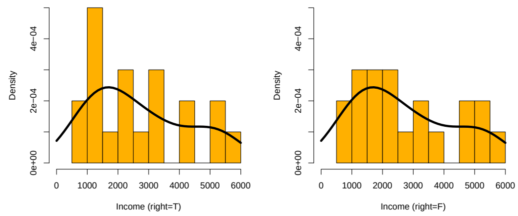
$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad , \qquad x \in \mathbb{R}$$

is the kernel density estimator with kernel K and smoothing parameter h.
For example $K(u) = \frac{3}{4}(1 - u^2)$ is the Epanechnikov kernel. The parameter h affects the width of the intervals within which observations are still considered in the kernel.

Notes

Example 2.7: Histogram vs. kernel density estimation

Wirtschafts- und Sozialstatistik
Universität Trier




```
x <- c(3500, 3200, 2100, 500, 1800, 2100, 5600, 4500, 1400,
1200, 1500, 2200, 3100, 1500, 2800, 1100, 5200, 4500,
5400, 800)
hist(x, probability=TRUE, breaks=11, xlim=c(0,6000), right=F,
col="#FFB000", main="", xlab="Income (right=F)",
ylab="Density")
lines(density(x, n=50, from=0, to=6000), lwd=4)
```


Notes

A two-dimensional frequency table

Wirtschafts- und Sozialstatistik
Universität Trier

Cat. of 1st variable \ Cat. of 2nd variable	1	$\cdots$	k	$\cdots$	r	Sum
1	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1r}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
j	n_{j1}	$\cdots$	n_{jk}	$\cdots$	n_{jr}	$n_{j\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$
m	n_{m1}	$\cdots$	n_{mk}	$\cdots$	n_{mr}	$n_{m\cdot}$
Sum	$n_{\cdot 1}$	$\cdots$	$n_{\cdot k}$	$\cdots$	$n_{\cdot r}$	n

Notes

Terminology of two-dimensional variables (1)

Wirtschafts- und Sozialstatistik
Universität Trier

- X First variable with x_j as j-th value ($j = 1, \ldots, m$)
- Y Second variable with y_k as k-th value ($k = 1, \ldots, r$)
- $n_{j\cdot}$ Absolute frequency of j-th value of variable X (marginal frequency)
- $n_{\cdot k}$ Absolute frequency of k-th value of variable Y (marginal frequency)
- n_{jk} Joint absolute frequency of j-th value of variable X and k-th value of variable Y
- $p_{j\cdot}$ Relative frequency of j-th value of variable X
- $p_{\cdot k}$ Relative frequency of k-th value of variable Y
- p_{jk} Joint relative frequency of j-th value of variable X and k-th value of variable Y

Notes

Terminology of two-dimensional variables (2)

Wirtschafts- und Sozialstatistik
Universität Trier

We have

$$\sum_{j=1}^{m} n_{j\cdot} = \sum_{k=1}^{r} n_{\cdot k} = \sum_{j=1}^{m}\sum_{k=1}^{r} n_{jk} = n$$

and

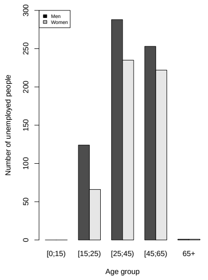
$$\sum_{j=1}^{m} p_{j\cdot} = \sum_{k=1}^{r} p_{\cdot k} = \sum_{j=1}^{m}\sum_{k=1}^{r} p_{jk} = 1\,.$$

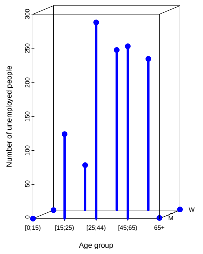
Notes

Example 2.8: Unemployment (see Example 2.1)

Wirtschafts- und Sozialstatistik
Universität Trier

Joint frequency distribution of unemployed people by age group and gender:





Notes

Notes

Notes