PV252

Tvorba uživatelských rozhraní a hodnocení použitelnosti

9. přednáška

ZPRACOVÁNÍ DAT: DESKRIPTIVNÍ STATISTIKA

Martin Dostál
Honeywell International - Aerospace Advanced Technology Europe
Masarykova Univerzita v Brně, Fakulta informatiky

Honeywell

POZOR, NAŠE ZPRACOVÁNÍ TÉTO PROBLEMATIKY JE S OHLEDEM NA PROSTOR ZNAČNĚ POVRCHNÍ. STATISTIKU JE TŘEBA STUDOVAT VE VĚTŠÍM DETAILU.

Data a znaky

- sběr dat
- znak (veličina)
  - kvantitativní
  - kvalitativní

Data a škály

- u kvalitativních dat možné hodnoty nazýváme kategoriemi
- nominální - data „pojmenovaná”, to však nevylučuje používání čísel - ovšem jako „jmen” či označení
- ordinální (pořadová) - data s uspořádáním
- intervalová (rozdílová) - ordinální data navíc s možností stanovit vzdálenosti mezi kategoriemi
- podílová - intervalová data navíc zachovávající podíl (tím pádem i násobení). Data jsou vždy kladná a mají tzv. absolutní nulu.

In nomine Patris et Filii et Spiritus Sancti

Příklady

- pohlaví (s možnými hodnotami mužské, ženské)
- barva očí (modrá, hnědá, černá)
- výsledek léčby (uzdraven, zemřel)
- národnost (česká, slovenská, polská, německá, ...)

- dosažené vzdělání (základní, střední, vysokoškolské),
- prospěch ve školním předmětu (výborně, velmi dobře, dobře, nevyhověl)
- stav pacienta (vyléčen, remise, recidiva)
- ohrožení povodní (stupně povodňové aktivity)
- hodnocení postojů ve škále (souhlasím, spíše souhlasím, spíše nesouhlasím, nesouhlasím)
- četnost výskytu (často, občas, zřídka, nikdy),
- chuť vína nebo jiné pochutiny podle degustátora atd.

- teplota ve stupních celsia
- datum

- váha, výška, věk
- rozměry, objem a hmotnost těles,
- koncentrace, kapacity,
- fyzikální vlastnosti materiálu, doba trvání nějakého děje,
- počet mikroorganismů ve vzorku vody,
- počet elementů ve vzorku krve atd.
- teplota v kelvinech (0 není možná)

Možné operace s daty dle použité škály

škála	operace
nominální	rovnost
ordinální	rovnost, uspořádání
intervalová	rovnost, uspořádání, součet/rozdíl
podílová	rovnost, uspořádání, součet/rozdíl, součin/podíl

Popis a prezentace kvalitativních dat

- absolutní četnost
- relativní četnost
- kumulativní četnost (absolutní/relativní)
- empirické rozdělení dat

Popis a prezentace kvantitativních dat

- obor hodnot - minimum, maximum
- kategorizace do intervalů
- histogram
- empirická distribuční funkce
- kvantily

Míry polohy a variability

- míry polohy
  - aritmetický průměr
  - geometrický průměr
  - median
  - modus
  - kvantily
- míry variability
  - výběrový rozptyl a směrodatná odchylka
  - kvartilové rozpětí
  - relativní kvartil

Centrální tendence a škály

- pro nominální data: modus
- pro ordinální data se průměr nehodí! Hodí se median a modus.
- pro intervalová data se hodí aritemtický průměr
- pro podílová data je vhodnější geometrický průměr (např. věk)

Geometrický průměr

Geometrický průměr *n* nezáporných čísel $x_1, x_2, \ldots, x_n$ je definován jako *n*-tá odmocnina jejich součinu:

$$G(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}.$$

- hodí se pro data s podílovou škálou
- hodí se, když má věcný význam součin znaků. Například při analýze znaků, které tvoří posloupnost a vznikají jako podíl dvou veličin (například tempo růstu)
- pro šipčatá data (špičatost - skewness) - asmyterické rozdělení

```
> data$growth
  [1] 23 30 24 24 27 21 25 23 22 23 20 24 27 25 24 26 52 25 19 19 18 19 19 19 19 19 18
18 18 19 18 21 20
 [34] 23 20 45 31 19 30 20 20 25 35 22 26 21 34 26 23 20 29 23 17 16 18 16 18 16 23 17
17 26 20 20 23 25
 [67] 23 30 24 20 21 28 31 20 20 21 21 54 19 20 20 19 19 20 21 20 20 20 20 20 20 19 19
20 20 20 20 20 19
[100] 35 19 20 20 19 19 20 20 19 20 24 27 48 21 25 22 17 32

> summary(data$growth)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   19.00   20.00   22.79   24.00   54.00

> geometric_mean(data$growth)
[1] 22.13008
```

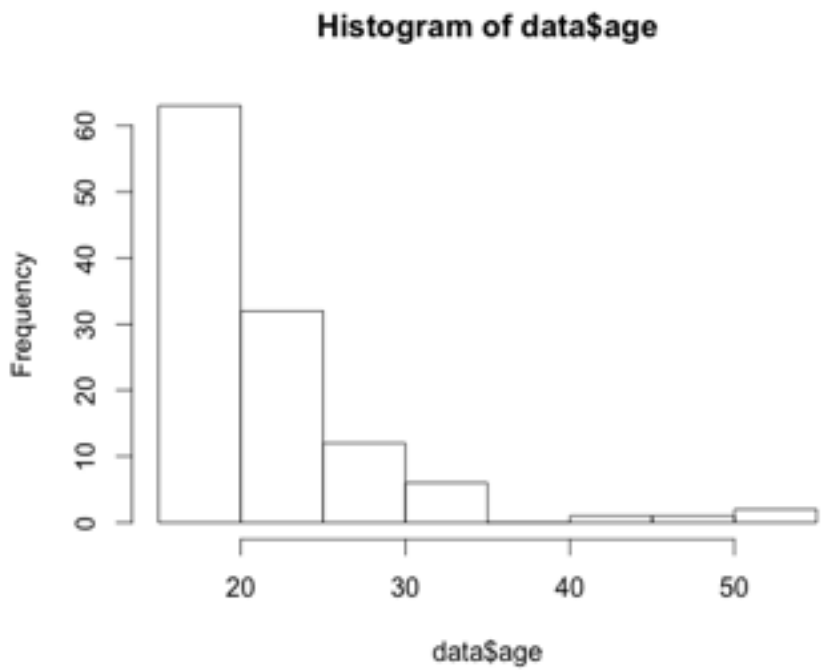
definice převzata z http://cs.wikipedia.org/wiki/Geometrický_průměr

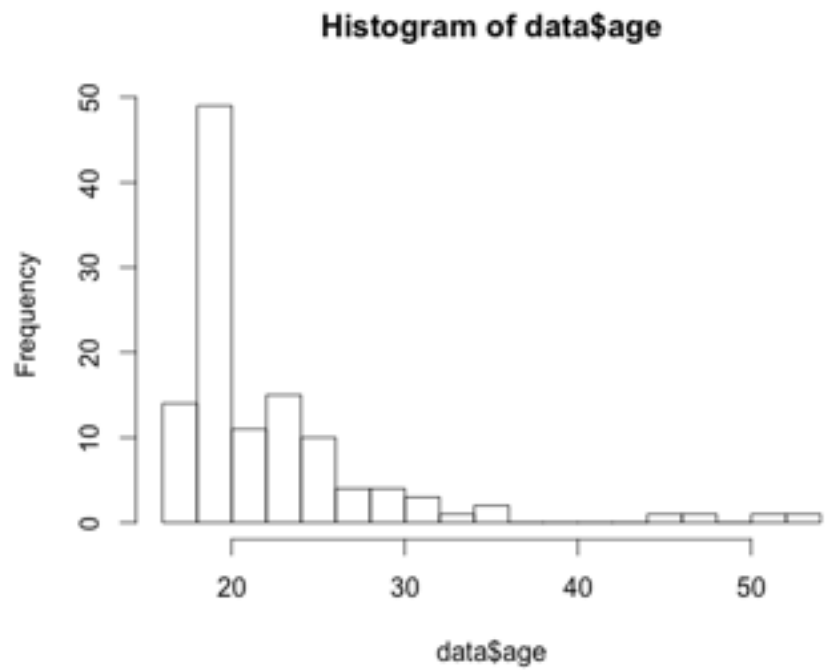
Grafická prezentace dat

- (pro nás) nejpoužívanější grafy
  - histogram
  - barplot
  - boxplot
- popisy os
- jednotky
- legenda

Histogram

- zachycuje absolutní četnost jednotlivých intervalů dělení
- toto dělení bývá nastavitelné



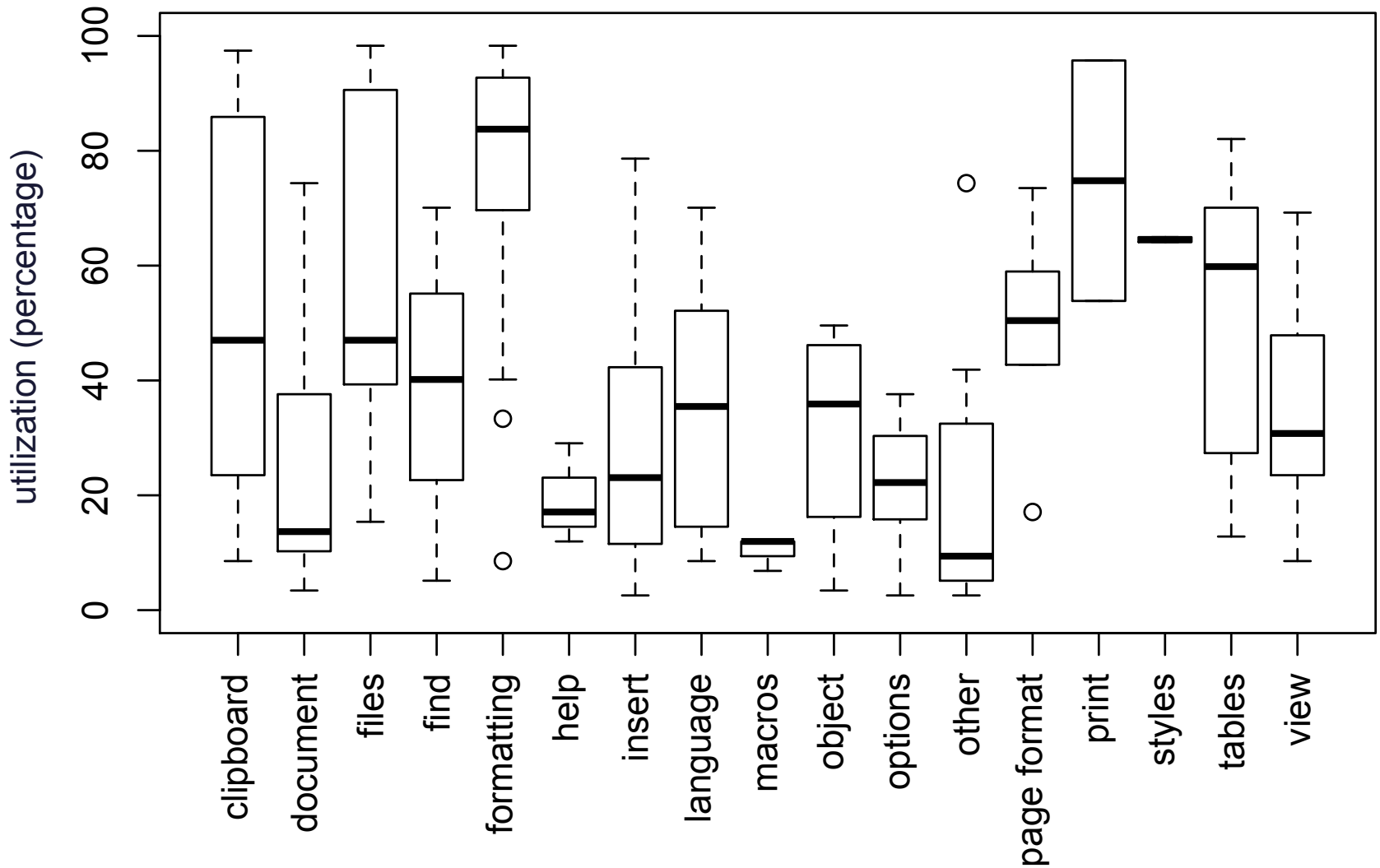



```
> hist(data$age)
> hist(data$age,breaks=20)
```

Boxplot

- česky „krabicový graf”
- vhodný pro zobrazení charakteristiky polohy a variability
- vyobrazuje
  - dolní a horní kvartil
  - median
  - horní a dolní vous (whiskers)
  - odlehlá pozorování (outliers)
- vyobrazení záleží na implementaci

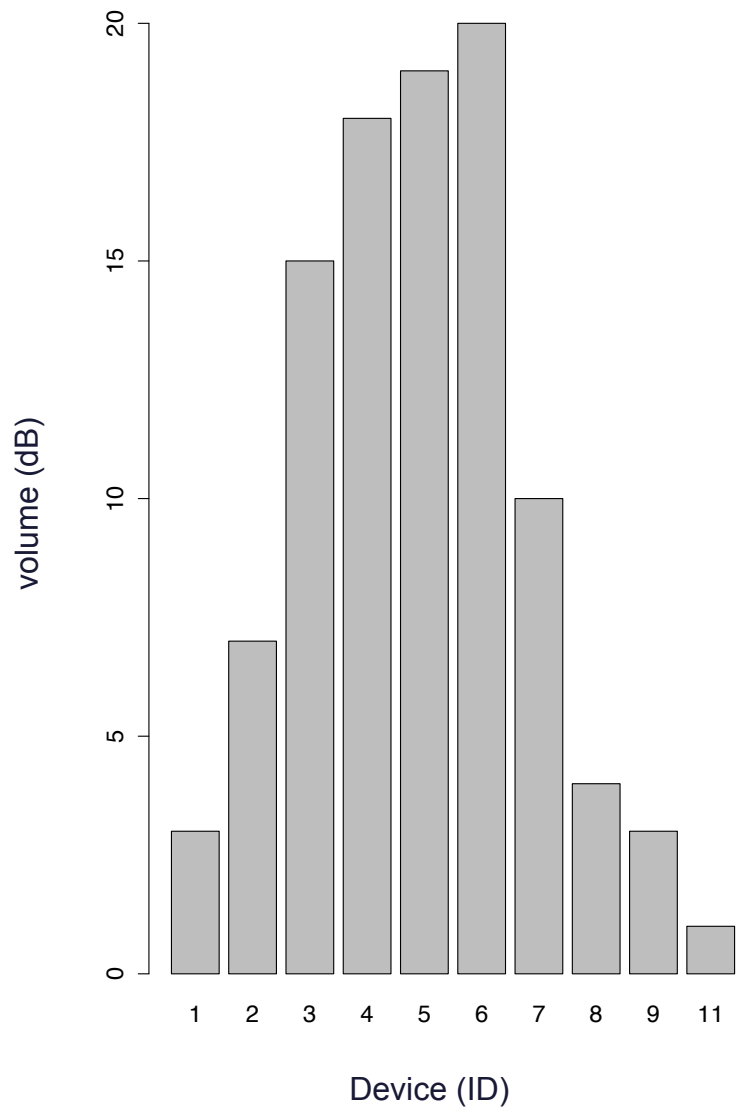
Boxplot


utilization (percentage)
0
20
40
60
80
100
clipboard
document
files
find
formatting
help
insert
language
macros
object
options
other
page format
print
styles
tables
view

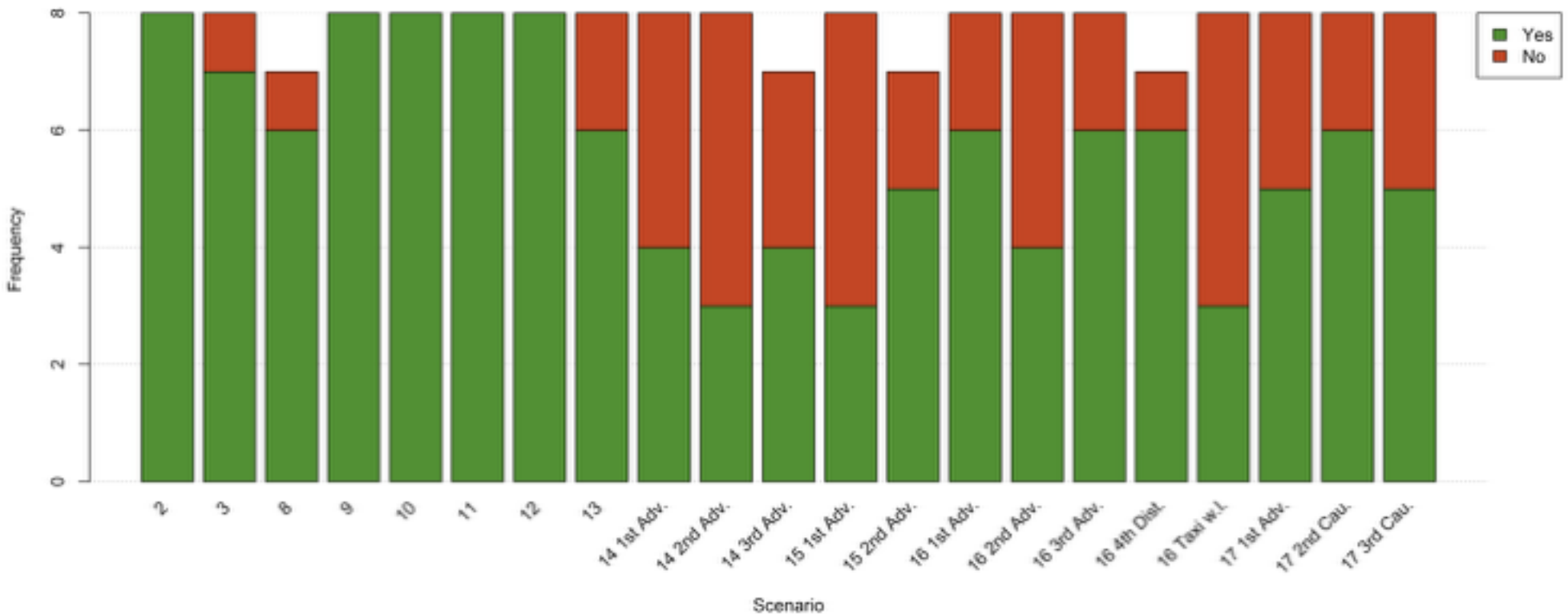
Barplot

- česky sloupcový graf
- vyobrazuje absolutní četnosti dat
- existuje též skupinová varianta
  - vyobrazení skupin vedle sebe (beside)
  - vyobrazení skupin na sobě (stacked)
- zobrazujeme-li tendenci, je vhodné doplnit o střední chybu (standart error) interval spolehlivosti (confidence interval)

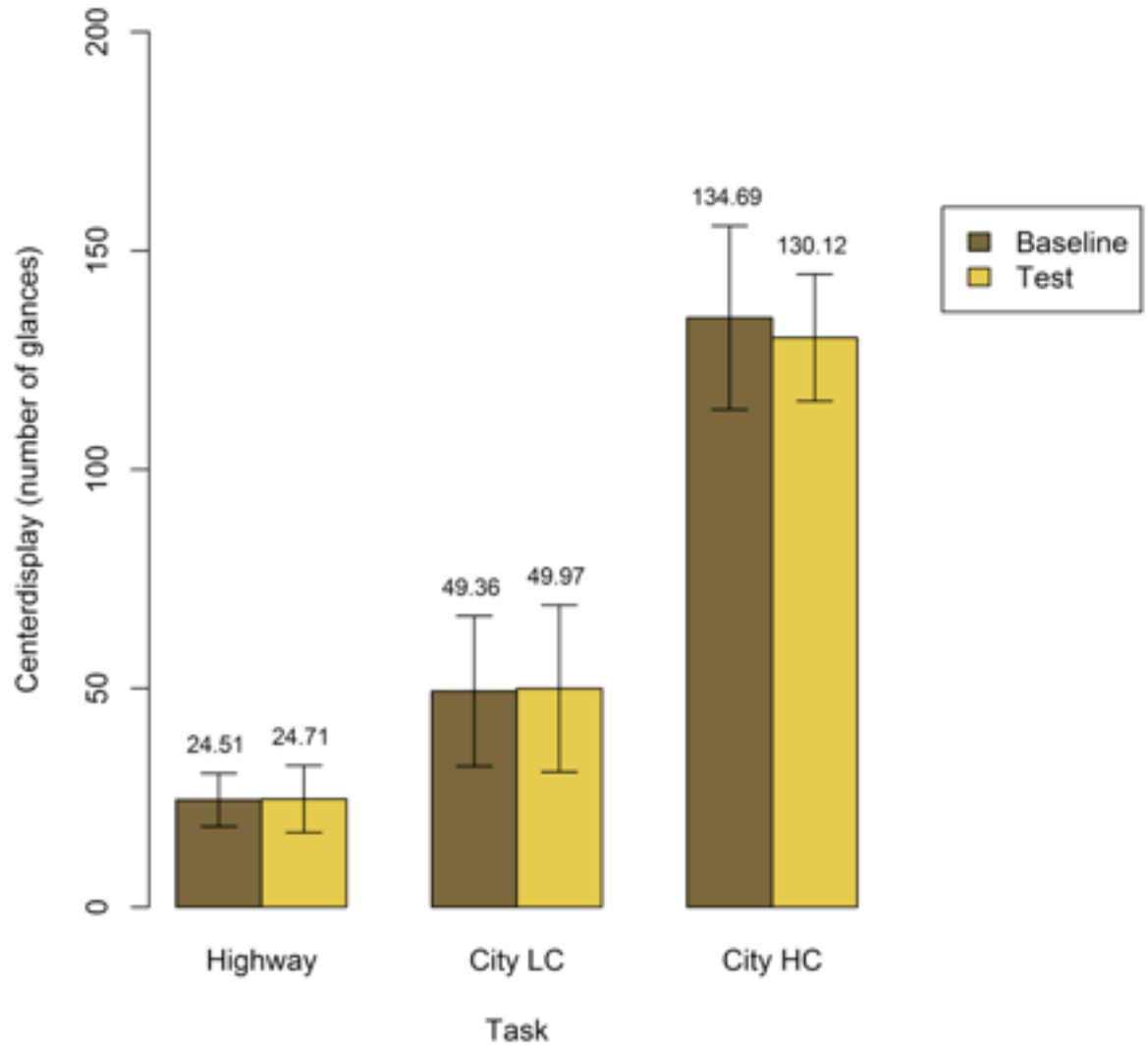
Barplot



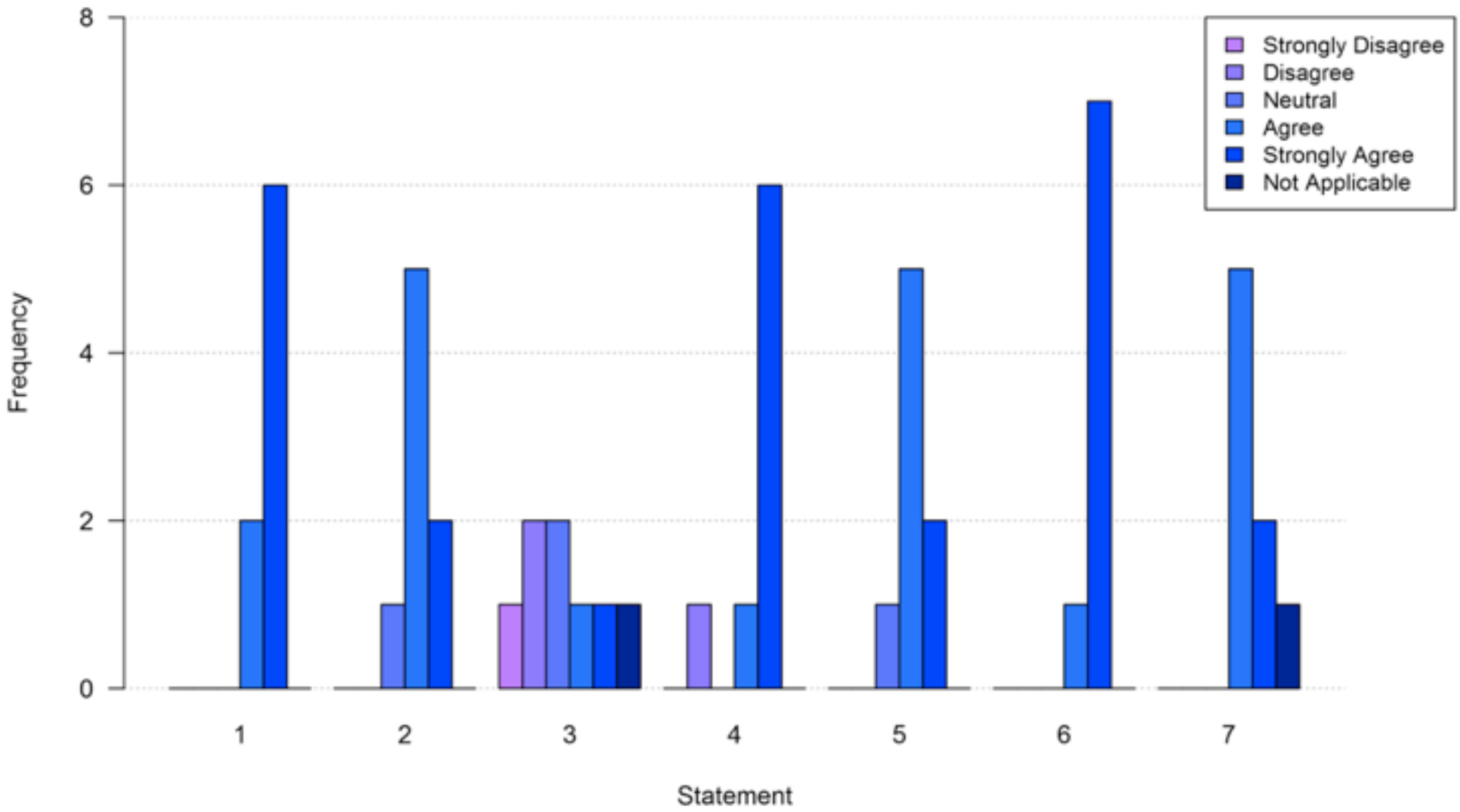
Barplot


Frequency
0
2
4
6
8
Yes
No
2
3
6
9
10
11
12
13
14 1st Adv.
14 2nd Adv.
14 3rd Adv.
15 1st Adv.
15 2nd Adv.
16 1st Adv.
16 2nd Adv.
16 3rd Adv.
16 4th Dist.
16 Taxi w.l.
17 1st Adv.
17 2nd Cau.
17 3rd Cau.
Scenario

Barplot s intervalem spolehlivosti



Barplot


Strongly Disagree
Disagree
Neutral
Agree
Strongly Agree
Not Applicable
Frequency
0
2
4
6
8
1
2
3
4
5
6
7
Statement

Reportování

- reportování deskriptivní statistiky
- podle APA
- důležité pro metodologickou správnost a srozumitelnost
- Průměr: M
- Median: Mdn
- Modus: Mode
- Směrodatná odchylka: SD
- Minimum: Min
- Maximum: Max
- Počet vzorků: N

Reportování

- **důsledně oddělit popis (reportování) od interpretace**
- hodnoty uvádíme v závorkách oddělené čárkou
- některé hodnoty můžeme však integrovat do textu pro lepší čitelnost
- zpravidla dvě desetinná místa, pro malé rozsahy i více
- deskriptivní statistika
  - průměr/medián/modus
  - směrodatná odchylka

Participants were 88 men and 100 women aged 16 to 34 years (men: *M* = 18.2, *SD* = 2.64; women: *M* = 21.4, *SD* = 2.12).

Četnosti v textu

- je vhodné vytvořit tabulku mapující adjektiva a četnosti
- potřebujeme-li uvést přesnou hodnotu, uvedeme ji v závorkách

The frequency of subjects responding is expressed as follows:

- "None of" (0 subjects),
- "Few", "Few of" (1, 2 subjects)
- "Some" (3, 4, 5 subjects), "Half of" (50% of responding subjects)
- Majority (> 50% of responding subjects)
- "Most" (6, 7 subjects)
- "All", [or wording not mentioned] (8 subjects)

In case of missing data the scale was adjusted. If the data cannot be collapsed the results were described as "Mixed".