WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Frequency distribution, cross tabulation, elementary hypothesis testing

Class 3: Marketing Research

Service and Digital Marketing
October 9, 2017

Chapter objectives

- Create descriptive statistics and graphs
- Calculate means and standard deviations of a distribution of observations
- Conduct χ^2 analyses and tests
- Understand how to use cross tables in practice and be able to interpret the results of different associated statistics
- R: working with vector and matrix objects (indexing, numerical operations), simple graph annotations

Cyanide and Happiness © Explosm.net

Descriptive statistics

- to obtain an initial idea of the dataset
- to perform data cleaning
- to determine the most important characteristics of different variables in a dataset
- different for nominal/ordinal (discrete) and metric (continuous) data = **levels of measurement**

Levels of measurement (1)

Discrete data

Nominal scale: categories or qualitative classifications

mathematical operations: $=, \neq$

e.g.: male, female

`R` **data type**: character, logical, factor

Ordinal scale: sorted categories

mathematical operations: $>, <, \geq, \leq$

e.g.: **likert scale**: completely agree, mostly agree, mostly disagree, completely disagree (subclass of **rating scale**, sometimes treated as "pseudo-meteric")

`R` **data type**: factor, numeric (integer)

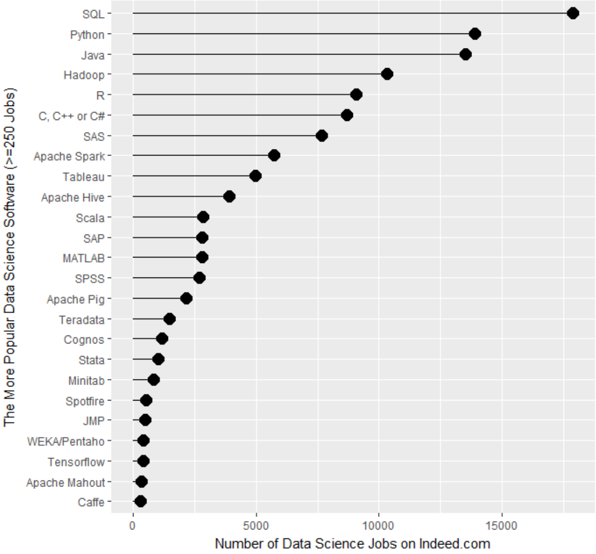
Levels of measurement (2)

Continuous data

Interval scale: scales with an arbitrary defined zero point
mathematical operations: $+, -$
e.g.: celsius scale, direction (measured in degrees from true or magnetic north), also sometimes rating scales (attitude and opinion scales)

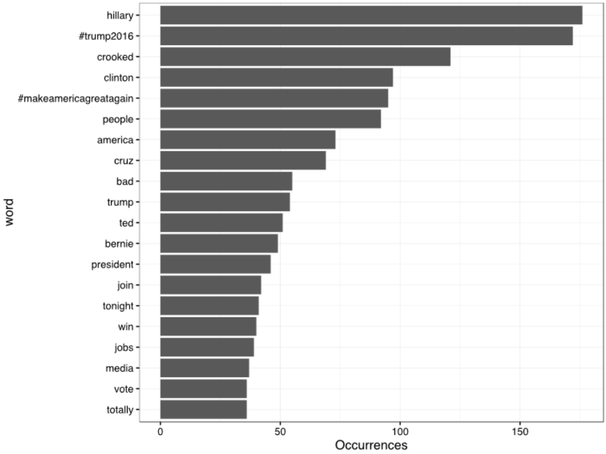
Ratio scale: possesses a meaningful zero value, most measurement in the physical sciences and engineering is done on ratio scales
mathematical operations: $\star, \div$
e.g.: kelvin scale, age, income, price, costs, sales revenue, sales volume, market share, ...
`R` **data type**: numeric (double)

Discrete variables: Example (1)



Discrete variables: Example (2)

What were the most common words in Trump's tweets?



Discrete variables: Frequency tables (1)

- to obtain a count of the number of responses associated with different values of **one** variable
- to indicate how scores of respondents are distributed over meaningful categories

Example: Trump's twitter behavior

5109 words obtained from tweets based on Trump's phones (during the presidential election campaign in 2016) had been categorized into 10 sentiments using the NRC Word-Emotion Association lexicon.

Research question: how is the word sentiment of tweets distributed?

Discrete variables: Frequency tables (2)

- obtain how responses are distributed over the range of possible values (number and percentages for each response category)

```
# generate the data
> name <- c("anger", "anticipation", "disgust", "fear",
            "joy", "negative", "positive", "sadness",
            "surprise", "trust")
> dat <- c(rep(name[1],490), rep(name[2],428), rep(name[3],304),
         rep(name[4],403), rep(name[5],355), rep(name[6],820),
         rep(name[7],967), rep(name[8],450), rep(name[9],266),
         rep(name[10],626))

# inspect data
> head(dat)
# obtain a frequency table
> table(dat)
```

Discrete variables: Frequency tables (3)

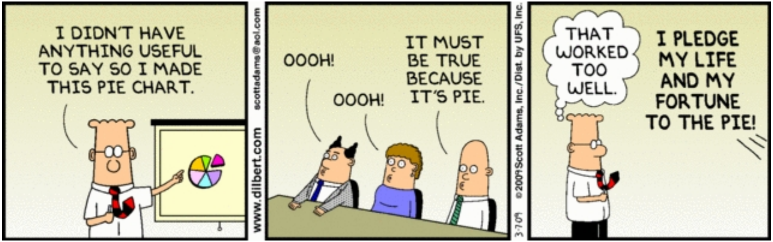
```
# relative values
> table(dat)/sum(table(dat))

     anger anticipation      disgust         fear          joy
0.09590918   0.08377373   0.05950284   0.07888041   0.06948522
  negative     positive      sadness     surprise        trust
0.16050108   0.18927383   0.08807986   0.05206498   0.12252887

# percentages
> table(dat)/sum(table(dat))*100

     anger anticipation      disgust         fear          joy
  9.590918     8.377373     5.950284     7.888041     6.948522
  negative     positive      sadness     surprise        trust
 16.050108    18.927383     8.807986     5.206498    12.252887

# total number of observations
> length(dat)
5109
# gives the same result (but taking NAs into account)
> sum(table(dat))
5109
```

Other possible research questions for frequency tables

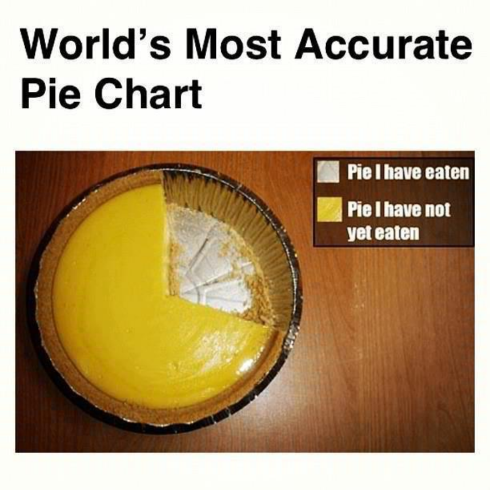
- What is the range of response values?
- What is the distribution of the responses?
  (e.g. highest? lowest?)
- How does it look like when responses are collapsed?
  (e.g. if there are few responses in some categories)
- Is there a substantial (e.g. neutral) response?
- How large is the missing data component and what effect does it have (on the results)?

Discrete variables: Bar charts (1)

- to display the results from a frequency table in a graph
- to depict the number of observations for every possible observed response (= visual representation of the data makes it easier to see patterns in it)



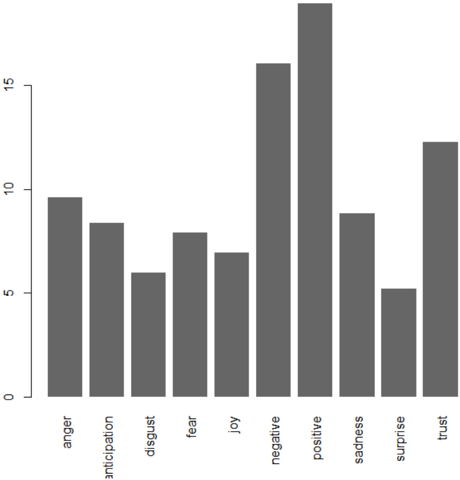
Discrete variables: Bar charts (2)



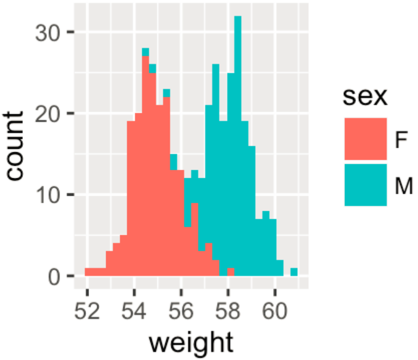
Discrete variables: Bar charts (3)

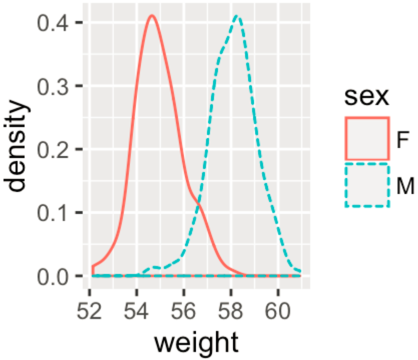
```
# frequencies
> barplot(table(dat))

# percentages
> barplot(table(dat)/sum(table(dat))*100)

# change the colors
> barplot(table(dat)/sum(table(dat))*100, col="grey40")

# delete the box borders
> barplot(table(dat)/sum(table(dat))*100, col="grey40", border=NA)

# rotate the bar labels
> barplot(table(dat)/sum(table(dat))*100, col="grey40", border=NA, las=3)

# hint: list of possible (named) colors available
> colors()
# for more annotations try
> ?barplot
```

Discrete variables: Bar charts (4)



Continuous variables: Example (1)





Continuous variables: Measures of location & dispersion (1)

- to determine the most important characteristics of non-nominal (i.e. **ordinal** or **continuous**) data
- to summarize the characteristics of a variable in one statistical indicator
- to provide an indication of the variability in a set of scores on a variable

- **mean** $\bar{x}$ (average):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

e.g.: $\bar{x} = (7 + 10 + 16 + 9 + 12 + 13 + 9 + 8 + 10 + 9)/10 = 10.3$

- **mode** (most frequent value):
e.g.: $mode = 9$ (appears 3 times)

- **median** $\tilde{x}$ (value in the middle):
e.g.: $7\ 8\ 9\ 9\ \underbrace{9\ 10}_{\tilde{x}=9+10}\ 10\ 12\ 13\ 16$, $\tilde{x} = 9.5$

- **variance** s^2 (mean squared deviation of the mean):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

e.g.: $((7-10.3)^2 + (10-10.3)^2 + \ldots + (9-10.3)^2)/(10-1) = 7.1222$

- **standard deviation** s (square root of the variance):
  e.g.: $s = \sqrt{7.1222} = 2.6687$

- **range** (spread of the data): $max - min$
  e.g.: $16 - 7 = 9$

- **interquartile distance** QD (difference between 25th and 75th percentile):

$$QD = Q_3 - Q_1$$

e.g.: $7\ 8\ \underbrace{9\ 9}_{25\%}\ \underbrace{9\ 10}_{\tilde{x}=50\%}\ \underbrace{10\ 12}_{75\%}\ 13\ 16$, QD = 12 - 9 = 3

Continuous variables: Measures of location & dispersion (2)

```
# generate some normal distributed fake weight data
> set.seed(1234)
> female <- rnorm(200, 55)
> male <- rnorm(200, 58)
> wdata <- data.frame(female=female, male=male)

# inspect the data
> head(wdata)

# summary statistics for the two variables
> summary(wdata)

     female             male      
 Min.   :52.14   Min.   :54.60  
 1st Qu.:54.23   1st Qu.:57.42  
 Median :54.83   Median :58.13  
 Mean   :54.94   Mean   :58.07  
 3rd Qu.:55.55   3rd Qu.:58.70  
 Max.   :58.04   Max.   :60.92  
```

Continuous variables: Measures of location & dispersion (3)

Obtain additional information

variance: mean squared deviation

std. dev.: square root of the mean squared deviation from the mean

range: spread of data (difference between lowest and highest value)

median: value in the middle

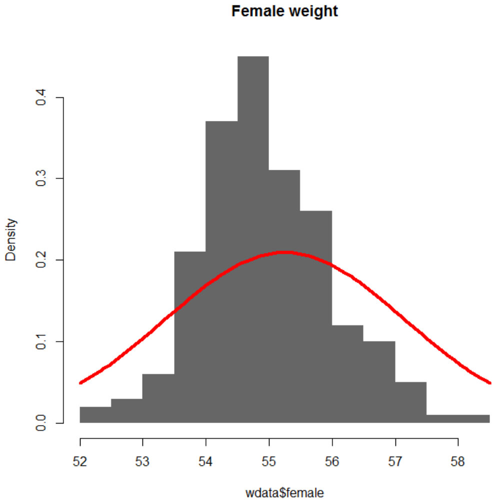
```
> var(wdata$female)
[1] 1.041759
> sd(wdata$female)
[1] 1.020666
> range(wdata$female)
[1] 52.14424 58.04377

> median(wdata$female)
[1] 54.82811
> mean(wdata$female)
[1] 54.94224
```

Continuous variables: Histogramm (1)

- display the distribution of a **continuous** variable by a number of **created** groups (continuous = here: nearly all observations have a different value)

```
# make a histogramm of the female weight
> hist(wdata$female, main="Female weight")

# change the color
> hist(wdata$female, main="Female weight", col="grey40")

# delete the bar border
> hist(wdata$female, main="Female weight", col="grey40", border=NA)

# add the normal distribution curve to the histogramm (set frequency to false)
> hist(wdata$female, main="Female weight", col="grey40", border=NA, freq=FALSE)
> x <- wdata$female
> curve(dnorm(x, mean(x), sd(x)), col = "red", lwd = 4, add = TRUE)
```

Continuous variables: Histogramm (2)



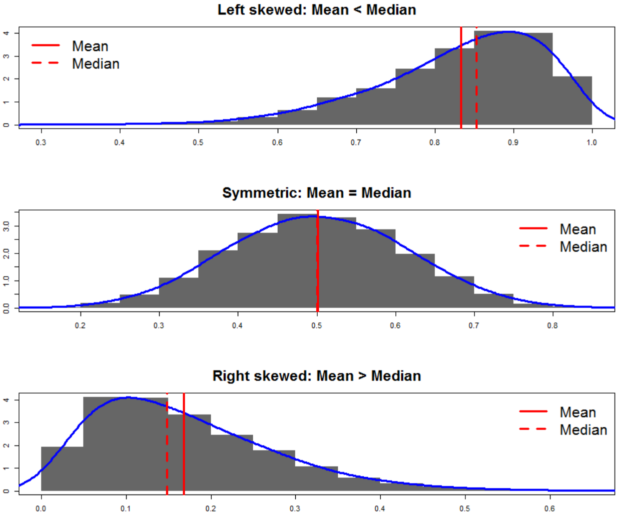
Summary descriptive statistics (1)

	nominal	ordinal	ratio (interval)
$\bar{x}$			$\checkmark$
$\tilde{x}$		$\checkmark$	$\checkmark$
mode	$\checkmark$	$\checkmark$	$\checkmark$
s^2, s			$\checkmark$
QD		$(\checkmark)$	$\checkmark$
range	$\checkmark$	$\checkmark$	$\checkmark$
Barplot	$\checkmark$	$\checkmark$	
Histogramm			$\checkmark$

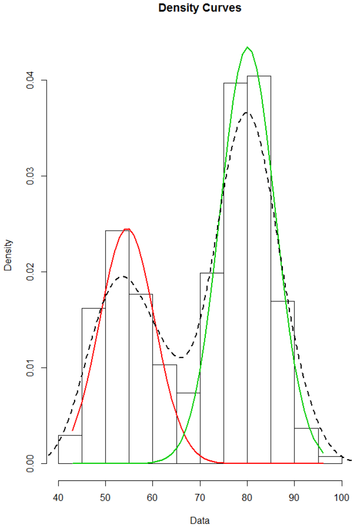
Note:

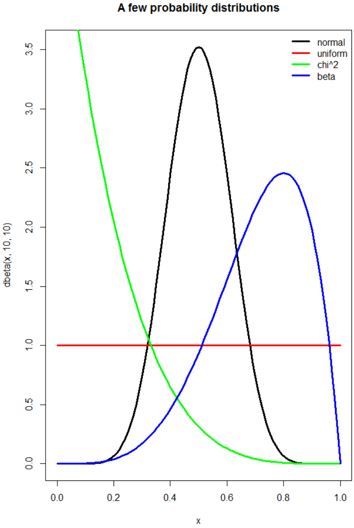
$\bar{x}$ and s^2 are only meaningful if the data is **symmetrically** distributed and **single peaked**!

Summary descriptive statistics (2)



Summary descriptive statistics (3)





See learning platform!

Submission deadline: 15. 10. at 23:00 pm

(via the learning platform `www.learn.wu.ac.at`)

Oral presentation of solutions on Monday!

(Recap: random selection of four students to present their solution).