ADVANCED STATISTICAL MODELING

Version 1, 2018

Jacolien van Rij [j.c.van.rij@rug.nl]
Hermine Berberyan, and Stefan Huijser

GOAL

Master HMC Cognitive Science

LEARNING TO ANALYZE DATA FROM HUMAN PARTICIPANTS USING REGRESSION TECHNIQUES IN R

-> open source
-> transparent
-> flexible

"ANALYZE DATA"

- **Statistics:**

> Statistics investigates and develops specific methods for evaluating hypotheses in the light of empirical facts. (Romein, 2017)

  - **Hypotheses:** general statements about population
  - **Data:** observations / measurements from experimental study
    - sample of population

"ANALYZE DATA"

- **Statistics:**

> Statistics investigates and develops specific methods for evaluating hypotheses in the light of empirical facts. (Romein, 2017)

- Two types:
  - Descriptive statistics
  - Inferential statistics

DESCRIPTIVE STATISTICS

- Describing data
  - Summarizing
  - No inferencing
- Necessary prerequisite for inferencing statistics
  - Data cleaning & preparation for analyses
- Driven by hypotheses
  - but also exploratory

INFERENTIAL STATISTICS

- Evaluation of statistical hypotheses based on the sample data
  - drawing conclusions about larger group (population) based on the sample data

PRACTICAL THINGS

- Two meetings per week:
  - One hour lecture / interactive discussion
  - One hour lab session
    - Hands-on tutorials using R / R Studio
- 3 homework assignments, each 5% of grade
  - Important to wrap up the subparts
  - Not submitted or too late: 0 points
- Final exam: written exam

TOPICS

- Linear regression (LM)
- Linear mixed-effects regression (LME)
- Generalized Additive Mixed Modeling (GAMM)
- Validity of analysis

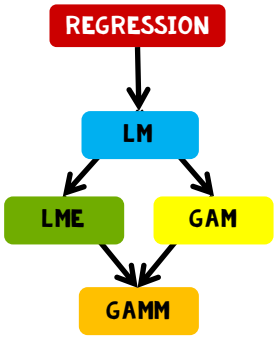
Schedule

Tentative schedule, might be updated during the course:

Week 1 (46): Linear model (LM)			
Tue	Nov 13, 2018	Introduction to the course Descriptive statistics and visualization	
Wed	Nov 14, 2018	Introduction linear model Basics linear regression	
Week 2 (47)			
Tue	Nov 20, 2018	Interactions in linear regression	
Wed	Nov 21, 2018	Model criticism and statistical significance	
Deadline assignment 1: Linear model			
Week 3 (48): Linear Mixed-Effects Model (LME)			
Tue	Nov 27, 2018	Introduction linear mixed-effects modeling	
Wed	Nov 28, 2018	Model criticism and statistical significance	
Week 4 (49)			
Tue	Dec 4, 2018	Generalized linear (mixed-effects) model	
Wed	Dec 5, 2018	Interpretation and reporting mixed-effects models	
Deadline assignment 2: Linear Mixed-Effects Model			
Week 5 (50): Generalized Additive Modeling (GAM)			
Tue	Dec 11, 2018	Introduction generalized additive modelling	
Wed	Dec 12, 2018	Nonlinear interactions	
Week 6 (51): Generalized Additive Mixed Modeling (GAMM)			
Tue	Dec 18, 2018	Parametric random effects in GAMMs	
Wed	Dec 19, 2018	Nonlinear random effects in GAMMs	
Deadline assignment 3: Generalized Additive Mixed Modeling			
Week 7 (2): Validity of statistical analysis			
Tue	Jan 8, 2019	GAMMs for modeling time series data	
Wed	Jan 9, 2019	Interpretation and reporting GAMMs	
Week 8 (3): Exam preparation			
Practice exam (obligatory to participate in discussion of practice exam)			
Tue	Jan 15, 2019	TBA	
Wed	Jan 16, 2019	Discussion practice exam	
Week 9 (4)			
Thu	Jan 24, 2019	Exam	

TOPICS

- Relation between the different regression method:



Today's topic

DESCRIPTIVE STATISTICS

DESCRIBING DATA

- Measures:
  - central tendency
  - variability
  - distribution

CENTRAL TENDENCY

- mean: $X = \frac{1}{N}\sum_{i=1}^{N} X_i$

```
mean(x, na.rm=TRUE)
```

- median: $\text{median} = \begin{cases} \text{odd}(N): & X_{(N+1)/2} \\ \text{even}(N): & \frac{X_{N/2}+X_{(N/2)+1}}{2} \end{cases}$

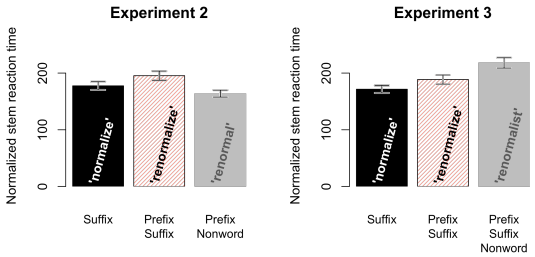
```
median(x, na.rm=TRUE)
```

- mode: most frequenty occurring elemen

```
tail(sort(table(x)),1)
```

AGGREGATION IN R

- Averages per group



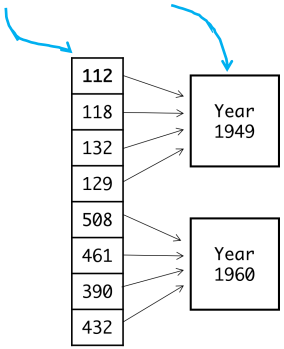
 (Feldman, Dale, van Rij, submitted)

AGGREGATION IN R

- Averages per group
- `tapply`, `aggregate`, `ddply` (package `plyr`)

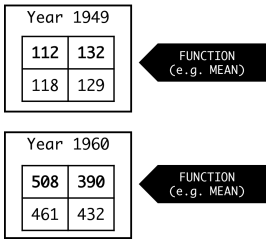
AGGREGATION IN R

`tapply( dat$Y , list(dat$Year), mean )`



AGGREGATION IN R

`tapply( dat$Y , list(dat$Year), mean )`



VARIABILITY

- St. deviance: $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - X)^2}$

  sd(x, na.rm=TRUE)
- Variance: var(x, na.rm=TRUE)

VARIABILITY

- Standard error of the mean (SE)
  - $s_X = \frac{s}{\sqrt{N}}$
  - Estimate of how far the **sample mean** is likely to be from the **population mean**
- Standard deviation of the sample:
  - Estimate of **variation within the sample**, how much the observations differ from the sample mean

VARIABILITY

- range: `range(x, na.rm=TRUE)`
- IQR: `IQR(x, na.rm=TRUE)`

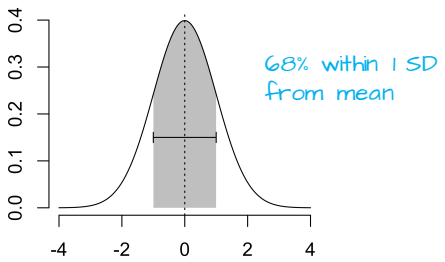
  `quantile(x, probs=c(.25,.75))`
- mean absolute deviation: $\frac{1}{N}\sum_{i=1}^{N}|X_i - X|$

  `mean( abs(x - mean(x)) )`
- median absolute deviation (MAD):

  `mad(x, constant=1)`

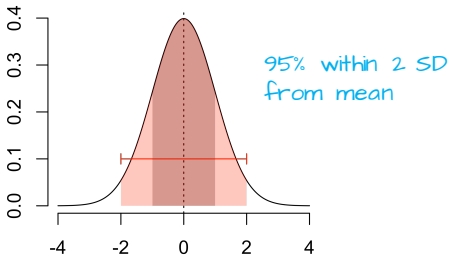
DISTRIBUTION

- Gaussian distribution: $X \sim N(\mu = 0, \sigma = 1)$



DISTRIBUTION

- Gaussian distribution: $X \sim N(\mu = 0, \sigma = 1)$

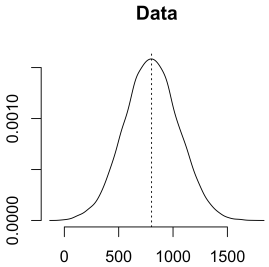


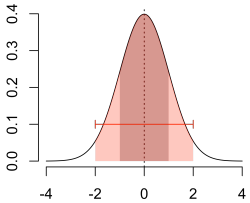
DISTRIBUTION

- Gaussian distribution: $X \sim N(\mu = 0, \sigma = 1)$

Z-SCORES

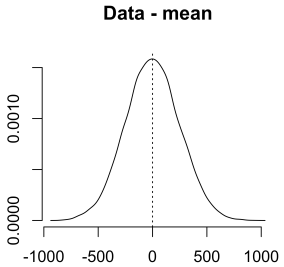
- Z-scores: $z_i = \frac{X_i - X}{s}$

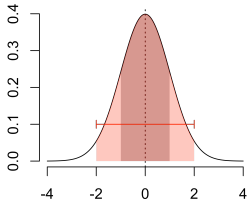




Z-SCORES

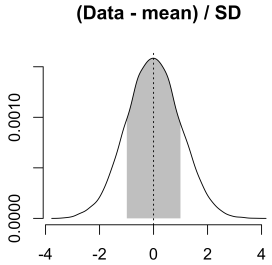
- Z-scores: $z_i = \frac{X_i - X}{s}$

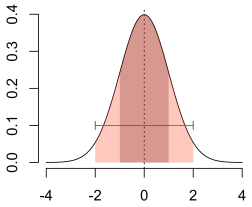




Z-SCORES

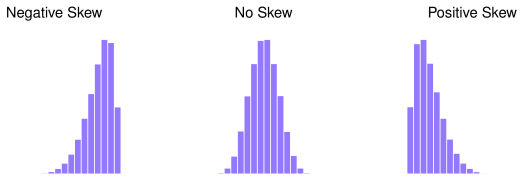
- Z-scores: $z_i = \frac{X_i - X}{s}$





DISTRIBUTION

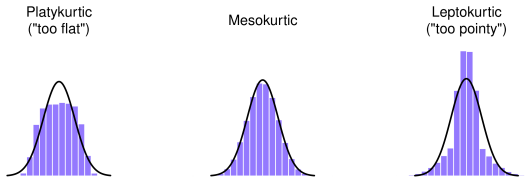
- Deviations from normal distribution



 (Navarro, 2017)

DISTRIBUTION

- Deviations from normal distribution



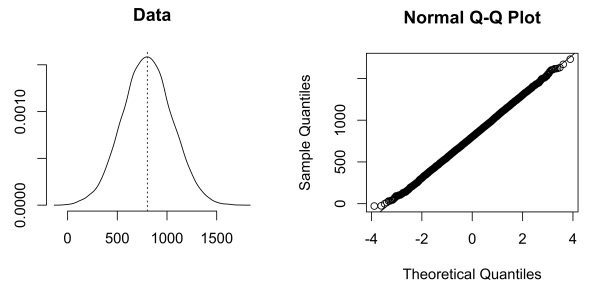
 (Navarro, 2017)

DISTRIBUTION

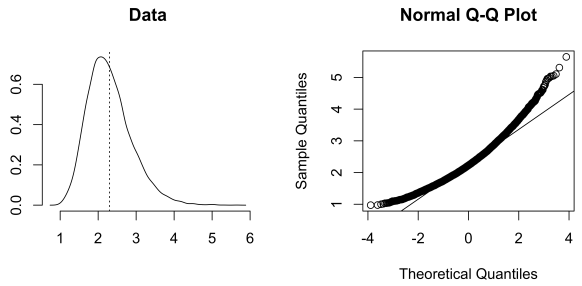
- histogram: `hist(x)`
- box plot: `boxplot(x)`
- density plot: `density(x)`
- QQ-plot: `qqnorm(x); qqline(x)`

 (Sprenger, la Roi, van Rij, submitted)

QQ-PLOT



QQ-PLOT



BUT BEFORE STARTING...

- Reading data into R
- Structure of data:
  - wide or long table?
  - coding of conditions
- Cleaning data