Part 5.
Correlation analysis

Natalia Levshina © 2017

University of Mainz, June 2017

Outline

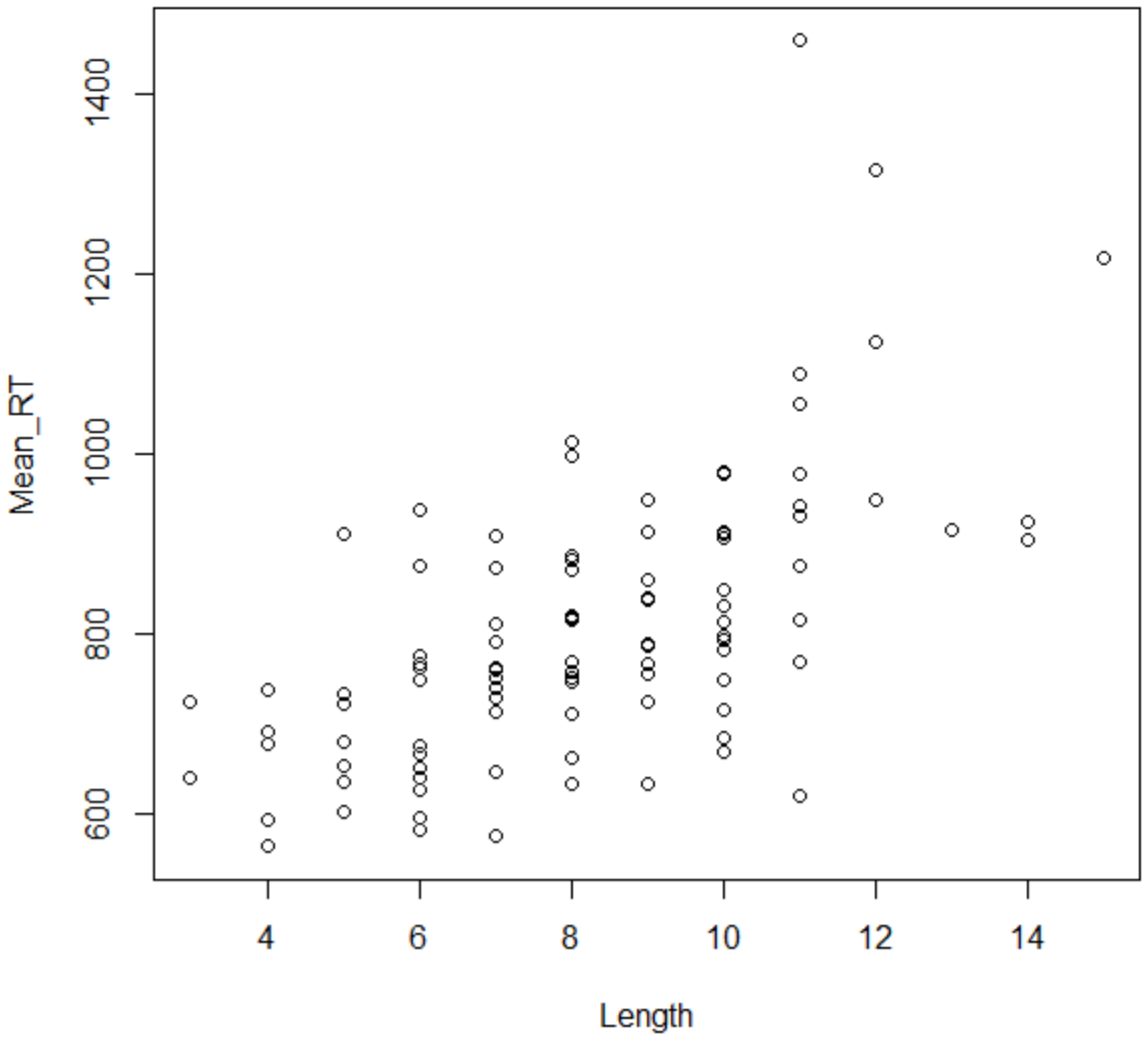
1. Word length and reaction times
    - Visualization of correlation
    - Pearson's correlation coefficient *r*
2. Acquisition of grammar and lexicon in L1
    - Visualization of correlation
    - Spearman's *rho*
    - Kendall's *tau*
3. Correlation and causation

Word length and reaction times

- Hypothesis: there is some correlation between word length and mean reaction times in a lexical decision task.
- This is a non-directional alternative hypothesis.
- Null hypothesis: no correlation.

Scatter plot

```
> library(Rling)
> data(ldt)
> attach(ldt)
> plot(Length, Mean_RT)
```


Mean_RT
600
800
1000
1200
1400
4
6
8
10
12
14
Length

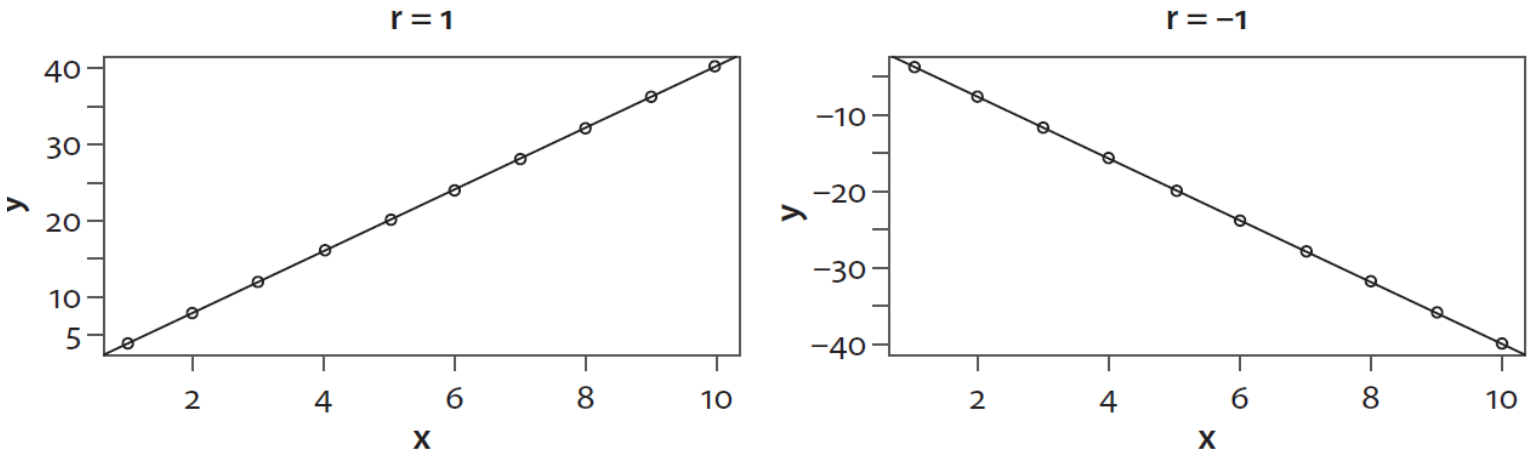
Correlation

- Is a relationship between two numeric variables
  - Positive: as X increases, Y increases, too. E.g. the more beers you drink before a psycholinguistic experiment, the longer your reaction times will be.
  - Negative, or inverse: as X increases, Y decreases. E.g. the more frequent a word, the shorter it is (Zipf's Law of Abbreviation).

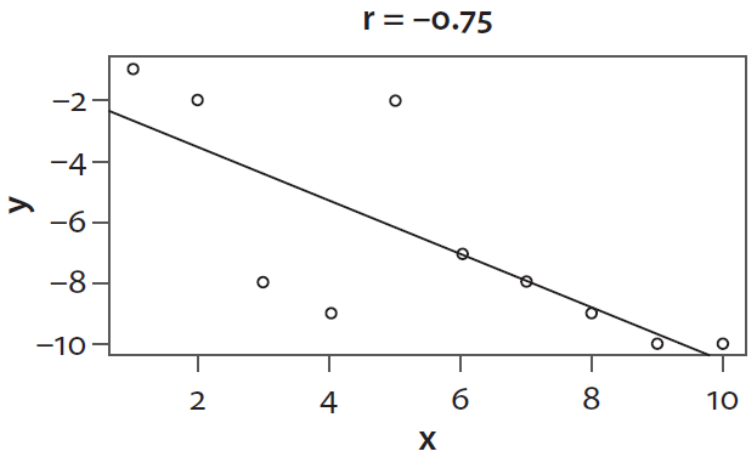
Correlation coefficient

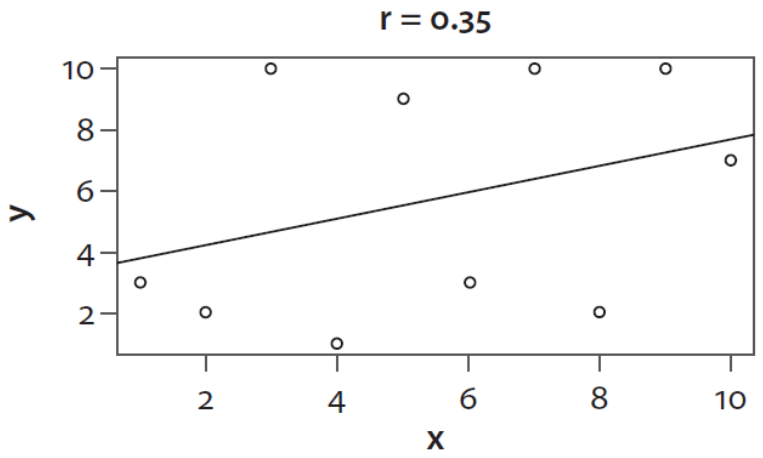
- A statistic from -1 to 1 which shows the direction and strength of a correlation.
  - Negative correlation: from -1 to 0
  - Positive correlation: from 0 to 1
  - No correlation: 0

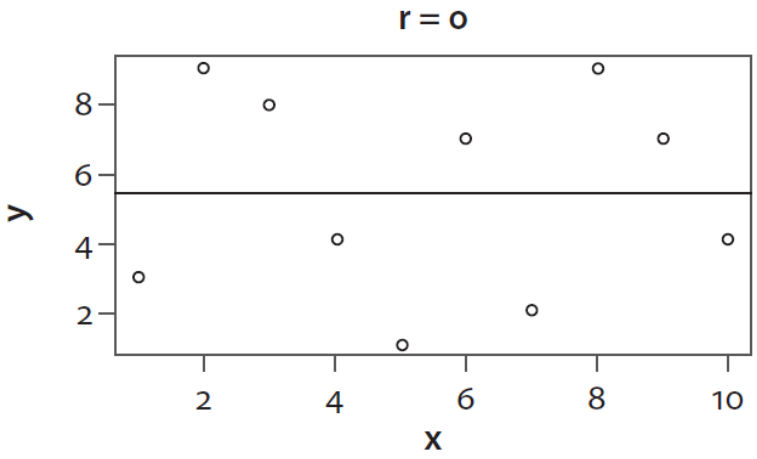
Perfect positive and negative correlations



Strong, weak and zero correlation





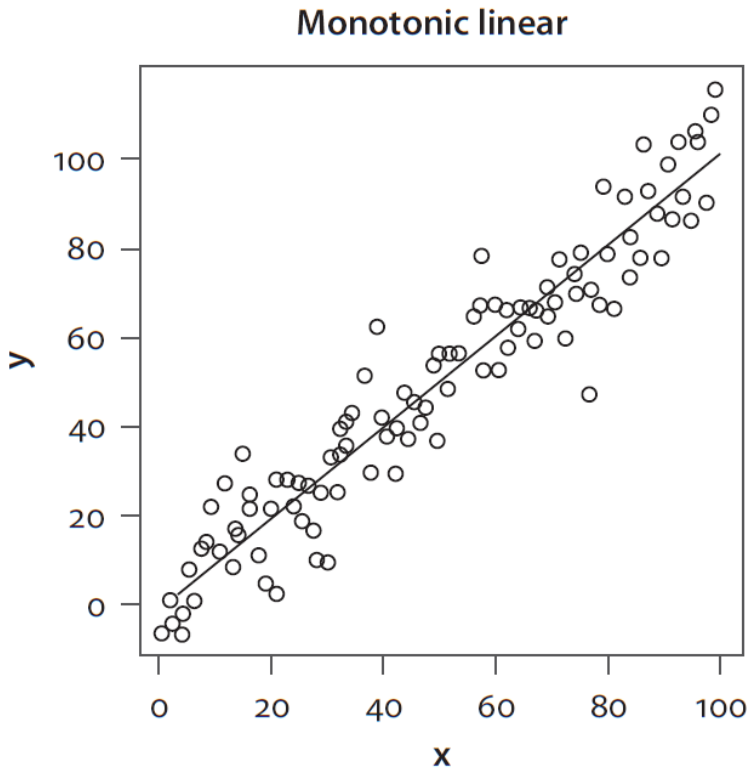


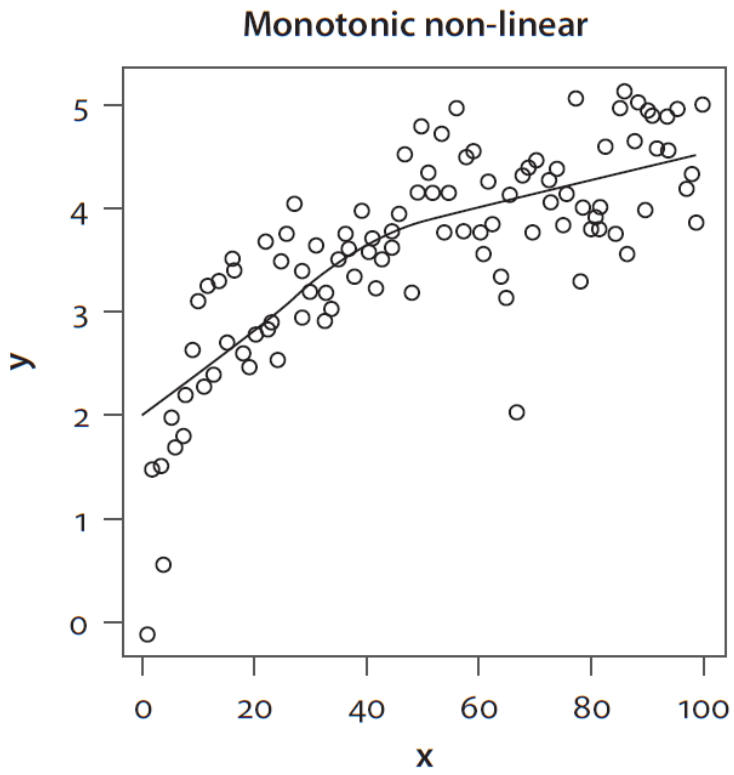
Exercise

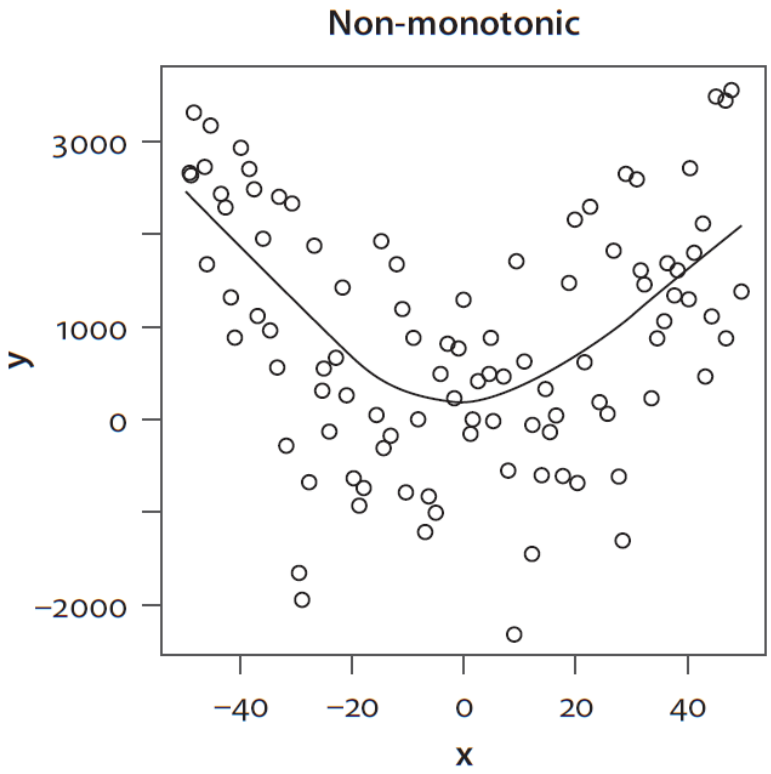
- Can you think of one example of a positive correlation, one example of a negative correlation, and one example of no correlation between two variables?

Types of relationships

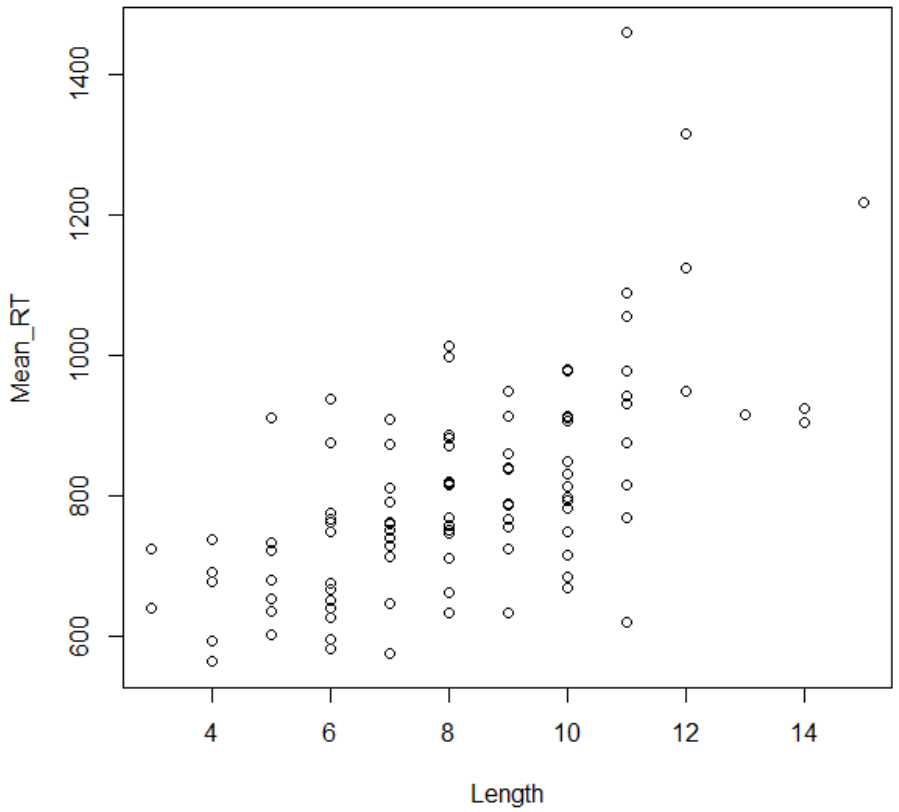
- Monotonic linear
  - Use Pearson's r
- Monotonic non-linear
  - Use Spearman's rho or Kendall's tau
- Non-monotonic
  - A more sophisticated method is needed


Monotonic linear
y
x
0
20
40
60
80
100


Monotonic non-linear
y
x
0
1
2
3
4
5
0
20
40
60
80
100


Non-monotonic
3000
1000
0
−2000
y
−40
−20
0
20
40
x

Question: what kind of relationship?



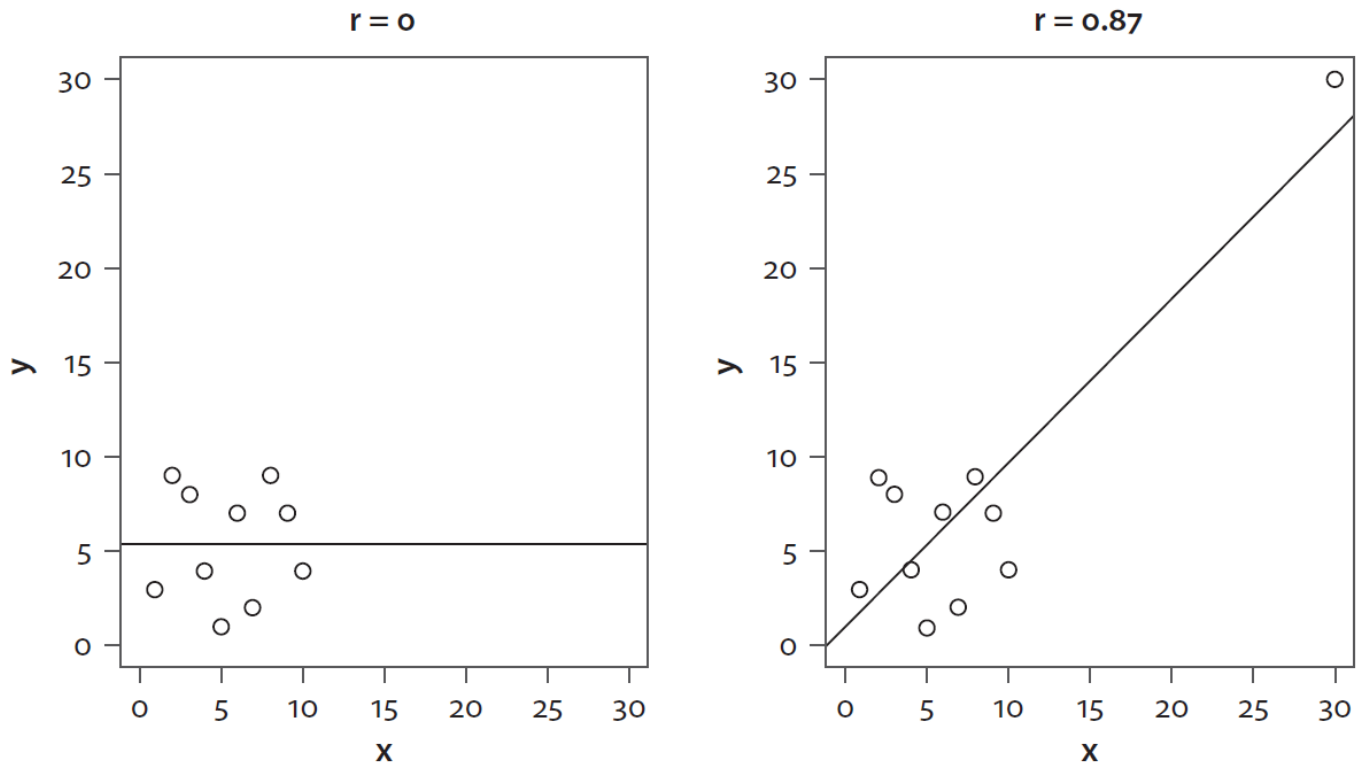
Pearson's *r* with all data

```
> cor.test(Length, Mean_RT) #by default, Pearson and two-tailed
        Pearson's product-moment correlation

data:  Length and Mean_RT
t = 7.7158, df = 98, p-value = 1.019e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4757761 0.7237704
sample estimates:
      cor
0.6147456
```

Beware of outliers

- Can distort the results:



Remove 3 old suspects

```
> Mean_RT_new <- Mean_RT[Mean_RT < 1200]
```

We also need to remove them from Length, so that the vectors are equally long:

```
> Length_new <- Length[Mean_RT < 1200]
```

Pearson's *r* without outliers

```
> cor.test(Length_new, Mean_RT_new)

        Pearson's product-moment correlation

data:  Length_new and Mean_RT_new
t = 7.0965, df = 95, p-value = 2.289e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4409167 0.7052541
sample estimates:
      cor
0.5886011
```

Interim conclusions

- The null hypothesis of no association can be rejected.
- The correlation is positive and significant.

Outline

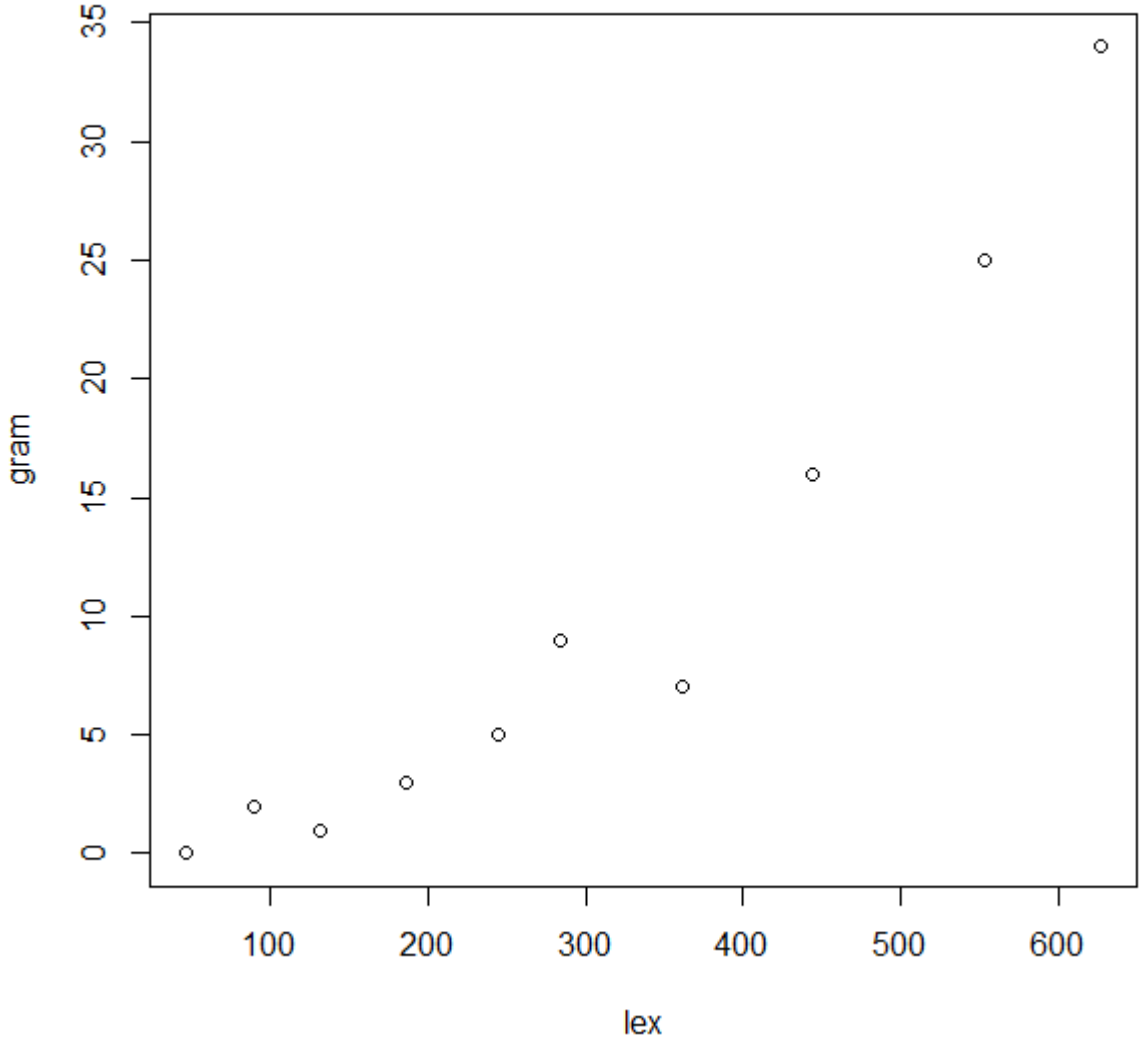
1. Word length and reaction times
    - Visualization of correlation
    - Pearson's correlation coefficient *r*
2. Acquisition of grammar and lexicon in L1
    - Visualization of correlation
    - Spearman's *rho*
    - Kendall's *tau*
3. Correlation and causation

Acquisition of grammar and lexicon

- Bates and Goodman (1997) investigated the relationships between vocabulary size and grammatical development of young children during the period from 16 to 30 months.
- Hypothesis: there is a positive correlation between vocabulary size and grammatical development.
  - Vocabulary size: the number of words learnt
  - Grammar: the total number of target constructions, from 0 to 37

Creating the data and a scatter plot

```
> lex <- c(47, 89, 131, 186, 245, 284, 362, 444, 553, 627)
> gram <- c(0, 2, 1, 3, 5, 9, 7, 16, 25, 34)
> plot(lex, gram)
```


gram
0
5
10
15
20
25
30
35
100
200
300
400
500
600
lex

Spearman’s rho

```
> cor.test(lex, gram, method = "spearman“, alternative = "greater") #one-tailed

        Spearman's rank correlation rho

data:  lex and gram
S = 4, p-value < 2.2e-16
alternative hypothesis: true rho is greater than 0
sample estimates:
      rho
0.9757576
```

Kendall’s tau

```
> cor.test(lex, gram, method = "kendall", alternative = "greater")

        Kendall's rank correlation tau

data:  lex and gram
T = 43, p-value = 1.488e-05
alternative hypothesis: true tau is greater than 0
sample estimates:
      tau
0.9111111
```

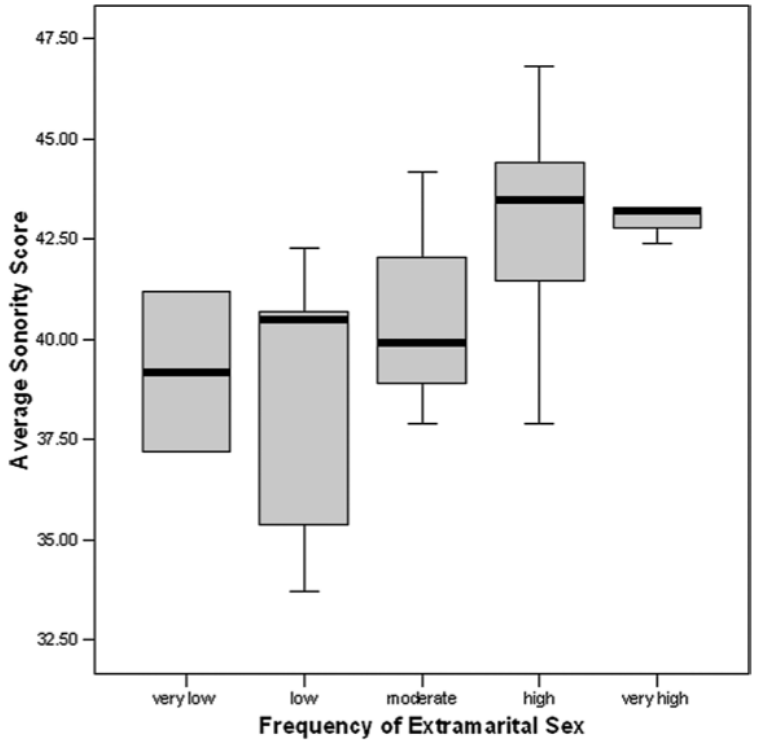
Interim conclusion

- The null hypothesis of no correlation can be safely rejected.
- There is a significant positive and very strong correlation between vocabulary size and grammatical development.

Outline

1. Word length and reaction times
    - Visualization of correlation
    - Pearson's correlation coefficient *r*
2. Acquisition of grammar and lexicon in L1
    - Visualization of correlation
    - Spearman's *rho*
    - Kendall's *tau*
3. Correlation and causation

Correlation is NOT causation

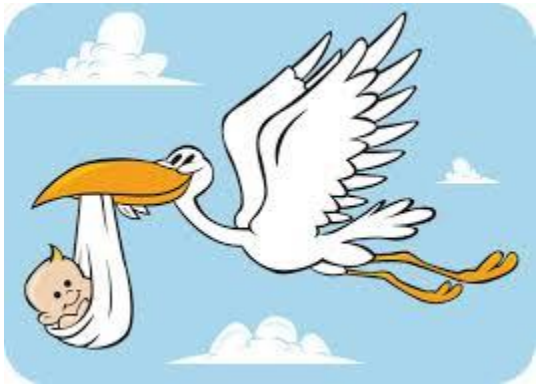


From Ember & Ember 2007

Spurious correlations

http://www.tylervigen.com/spurious-correlations

Theory of the Stork: a positive correlation between the number of storks and that of newborns -> do storks bring babies?!



Exercise

- Everett (2013) reported a positive correlation between elevation and the likelihood that a language contains ejective consonants:
  - “We suggest that ejective sounds might be facilitated at higher elevations due to the associated decrease in ambient air pressure, which reduces the physiological effort required for the compression of air in the pharyngeal cavity–a unique articulatory component of ejective sounds. In addition, we hypothesize that ejective sounds may help to mitigate rates of water vapor loss through exhaled air. ”
- Can you think of another way of explaining this correlation?

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0065275