TO JORN, CORINE, THERA AND TINEKE

DRAFT

Chapter 1

An introduction to R

```
> 6 / 3                  # division
[1] 2
> 2 ^ 3                  # power
[1] 8
> 9 ^ 0.5                # square root
[1] 3
```

The hash mark #

```
> nchar(c("antidisestablishmentarianism", "a"))
[1] 28  1
```

When applying `nchar()`

This contingency table can be recast as a table of proportions by dividing each cell in the table by the sum of all cells, the total number of observations in the data frame with inanimate themes. We obtain this sum with the help of the function

1.5 Session management

`R` stores the objects it created during a session in a file named `.RData`, and it keeps track of the commands issued in a file named `.Rhistory`

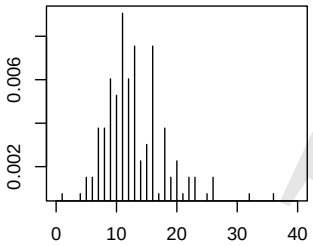
Chapter 2

```
> boxplot(exp(lexdec$RT))    # upper panel
> boxplot(lexdec$RT)         # lower panel
```

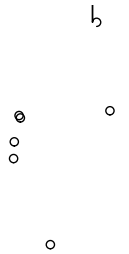
(For the upper panel, we use the exponential function `exp()` **to undo the logarithmic transformation of the reaction times in the data frame** `lexdec`

conditions? To answer this question, we use the function `dbinom()` **that we already introduced above. Given an observed value (its first argument), and given the parameters n and p (its second and third arguments), it returns the requested probability:**

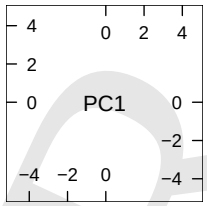
```
> dbinom(1, size = 1000000,  prob = 0.0000082)
[1] 0.002252102
```



```
> wonderland$alice = wonderland$word=="alice"
       word chunk alice
1    alice’s     1 FALSE
```



the verb). The column



Scatter Plot Matrix

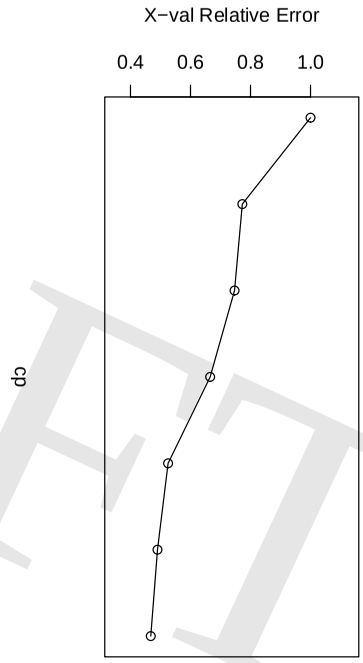
corpus-based computational semantics, latent semantic analysis [Landauer and Dumais, 1997].

In principal components analysis, the total variance is partitioned among the PCs. Therefore, the proportion of variance explained by a PC is given by that PC's variance divided by the summed variance of all PCs, as we saw above. In factor analysis, how-

```
> length(btr) = B
```

We now create 200 **bootstrap trees, sampling with replacement from the columns of our data matrix.**

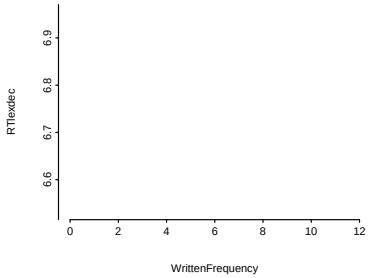
```
> for (i in 1:B) {
+  trB = nj(dist(papuan.mat[ ,sample(ncol(papuan.mat), replace = TRUE)],
+       method = "binary"))
```



when the effect of one predictor is codetermined by the value of another predictor. Fig-

Chapter 6

Regression modeling



```
> collin.fnc(items.pca$x, 1:4)$cnumber
[1] 1
```

Finally, we add these four principal components to our data, first for the young age group,

Frequency’’’

```
[1] 2419 2458 2844 2883 3628 3714 3815 3850 4381

$WrittenSpokenFrequencyRatio
[1] 1036 2400 2612 3320 3328 4148 4365

$AgeSubject
[1]  385 2097 2419 2458 2844 3628 3714 3815 3850 4381
```

```
> proportions = nCorrect/30
> proportions
[1] 0.9000000 1.0000000 1.0000000 1.0000000 0.8666667
[5] 0.9333333 1.0000000 0.9333333 0.8333333 0.9666667
```

noise is also fitted. Data points with extreme values due to noise are taken just as seriously as normal data points. Across experiments, it is unlik

measures at 40 **intervals with** 648 **tokens in each interval.**

```
> alice[1:4]
[1] "alice's"    "adventures" "in"         "wonderland"
> alice.growth = growth.fnc(text = alice, size = 648, nchunks = 40)
```

The output of `growth.fnc()` **is a growth object, and its contents can be inspected with** `head()` **or**

types with frequencies 1 **through** 5

Chapter 7

7.2.1 Mixed-effect models and Quasi-

The small p-value shows that we need to stay with the original, full model. Note that we

```
3       S3    long 546.25
4       S4    long 521.00
5       S5    long 569.75
6       S6    long 529.50
7       S7    long 490.00
8       S8    long 585.00
9       S1   short 556.50
10      S2   short 556.50
11      S3   short 579.25
12      S4   short 551.75
13      S5   short 594.25
14      S6   short 572.50
15      S7   short 535.75
16      S8   short 560.00
```

This high error rate carries over to the `F1+F2` procedure. As expected for small samples, the p-values for

Table 7.1: Type I error rate and power comparison for four regression models (`lmer`

In principle, we can include a random effect for `Speaker`

```
> selfPacedReadingHeid.lmer = lmer(RT ~ RTtoPrime + Condition +
+ (1|Subject) + (1|Word),  data = selfPacedReadingHeid)
> selfPacedReadingHeid.lmer
```

structure with the reading times of the words in the immediately preceding discourse. There is no way of doing so with traditional analyses requiring prior averaging over subjects and items.

```
> anova(beginningReaders.lmer4, beginningReaders.lmer3)
```

Appendix A

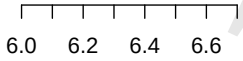
Solutions to the exercises

1.1

```
> spanishMeta
   Author YearOfBirth  TextName PubDate Nwords    FullName
1       C        1916 X14458gll    1983   2972        Cela
```

warlpiri.xtabs

CaseMarking

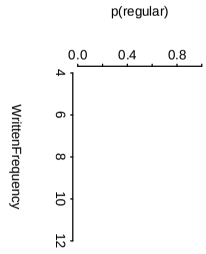


3.5

```
> plot(qpois(1:20 / 20, mean(countOfAlice)), quantile(countOfAlice,
```

```
> validate(etym.lrm, bw = T, B = 200)
...
         Frequencies of Numbers of Factors Retained

 2  3  4  5  6  7  8  9
 1  1 14 28 39 39 41 37
           index.orig     training       test     optimism
Dxy         0.621491185  0.639033808 0.549117995  0.089915814
R2          0.382722800  0.404198572 0.300891734  0.103306838
Intercept   0.000000000  0.000000000 0.004753967 -0.004753967
```



```
...
Population size: S = 810.356
...
Goodness-of-fit (multivariate chi-squared test):
         X2 df         p
```

Appendix B

FACTORS

`ordered()`	**create ordered factor**
`as.factor()`	**convert into factor**
`as.character()`	**convert factor into character vector**
`relevel()`	**select new reference level**
`[drop=TRUE]`	**drop unused factor levels**