Linguistic Data: Quantitative Analysis and Visualisation

Ilya Schurov, Olga Lyashevskaya, George Moroz, Alla Tambovtseva

26 January 2019

Hypothesis testing: example from lecture continued

So as to understand an example that will be discussed further, we should know something about matrices in R. A matrix is just a table of values, a set of vectors in R. Elements of a matrix are always of the same type (only numeric, only character, or only logical).

Let's create a matrix 2×3 consisting of zeroes:

```
# element, than number of rows and number of columns
matrix(0, nrow=2, ncol=3)
```

```
##      [,1] [,2] [,3]
## [1,]    0    0    0
## [2,]    0    0    0
```

Instead of a copying a single value, we can arrange a set of values and create a matrix. Let's arrange a vector of 12 values into matrix 3×4 in a way that values are arranged by rows: when a row is filled, values go to the next one.

```
v <- 1:12  # sequence of integers from 1 to 12
m <- matrix(v, n=3, ncol=4, byrow = TRUE)
m
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    5    6    7    8
## [3,]    9   10   11   12
```

Note: so as to understand the difference, you can skip the option `byrow=TRUE`:

```
# values go by columns
matrix(v, n=3, ncol=4)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

We can calculate the sums of every row in a matrix (marginal sums):

```
rowSums(m)
```

```
## [1] 10 26 42
```

Now let's recall how to create a sample with repeated values (with replacememt) of some values in R. We will create a sample of 0 and 1 of size 10.

```
sample(c(0, 1), 10, replace = TRUE)
```

```
##  [1] 1 1 1 1 1 1 1 0 0 0
```

Note: we have not set a seed, a starting point of the algorithm, so it is ok if you got a different sample of zeroes and ones.

Now we are ready to discuss an example of the experiment. Recall an experiment from the lecture: we toss a coin 10 times and repeat this sequence of tosses 10000 times (you can think of 10000 researchers who independently toss a coin 10 times). If we get a head, we write 1, if we get a tail, we write 0. Let's create a matrix that will contain the results of such an experiment:

```
tosses <- 10
samples <- 10000
dat <- matrix(sample(c(0, 1), tosses * samples, replace=TRUE), ncol=tosses, byrow=TRUE)
```

Comments: we created a matrix 10000×10, with 10000 rows and 10 columns that consists of 0 and 1. Each row in a matrix represents a sequence of 0 and 1, the results of tossing a coin 10 times.

```
head(dat)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    0    1    0    0    1    1    0    1    1     0
## [2,]    1    0    0    0    1    0    0    1    1     0
## [3,]    1    0    0    0    1    1    1    1    1     1
## [4,]    1    0    1    0    0    1    0    1    0     0
## [5,]    0    1    0    1    1    0    1    1    0     0
## [6,]    1    1    1    1    0    0    0    1    1     0
```

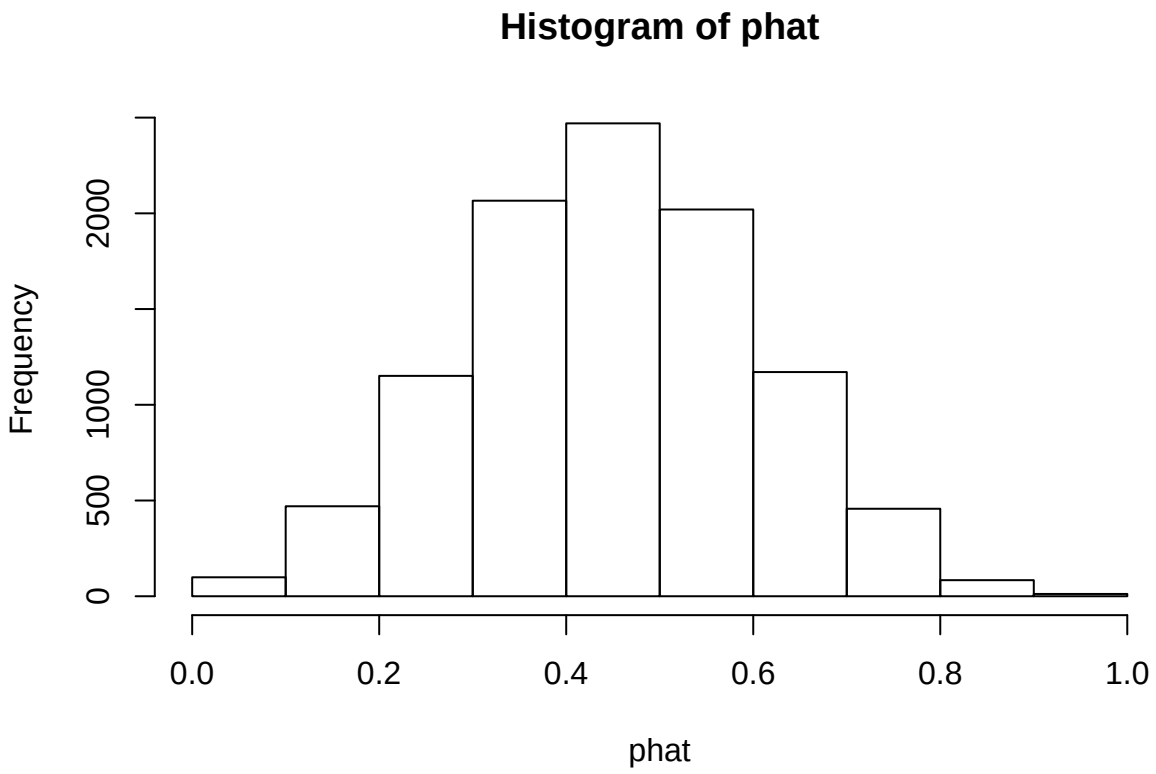
Let's calculate `phat`, proportions of heads in each series of tosses:

```
phat <- rowSums(dat) / tosses
head(phat)
```

```
## [1] 0.5 0.4 0.7 0.4 0.5 0.6
```

Let's plot a histogram of `phat`:

```
# xlim - limits for x axis
# we want to see all values from 0 to 1
hist(phat, breaks=tosses, xlim=c(0, 1))
```



As we can see, getting 10 heads (`phat=1`) or 10 tails (`phat=0`) is almost impossible while getting 3 or 5 heads is highly probable. Now let's proceed to formal tests.

We will test $H_0 : p = 0.5$. In other words, we will consider different results (different numbers of heads in 10 tosses) and decide whether a coin tossed was fair.

```
# 3 out of 10 - a fair coin?
binom.test(3, 10, p=0.5)
```

```
##
##  Exact binomial test
##
## data:  3 and 10
## number of successes = 3, number of trials = 10, p-value = 0.3438
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.06673951 0.65245285
## sample estimates:
## probability of success
##                    0.3
```

As we can see, p-value is greater than a significance level $\alpha = 0.05$, so we have no grounds to reject the null hypothesis based on the data given.

NB: Judging by results of a statistical test, we *cannot* decide whether our null hypothesis is true or not. The only thing we can say is that our data provide no grounds for rejecting the null hypothesis at a certain significance level α. For the same reason, we *never* say that we accept our null hypothesis. So,

$$\text{not rejecting } H_0 \neq \text{accepting } H_0 \neq H_0 \text{ is true}$$

Now make your own conclusions and test your intuition!

```r
# 2 out of 10 - a fair coin?
# 1 out of 10 - a fair coin?
binom.test(2, 10)
```

```
##
##  Exact binomial test
##
## data:  2 and 10
## number of successes = 2, number of trials = 10, p-value = 0.1094
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.02521073 0.55609546
## sample estimates:
## probability of success
##                    0.2
```

```r
binom.test(1, 10)
```

```
##
##  Exact binomial test
##
## data:  1 and 10
## number of successes = 1, number of trials = 10, p-value = 0.02148
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.002528579 0.445016117
## sample estimates:
## probability of success
##                    0.1
```

Note: the probability `p=0.5` is set by default in R, so we can skip this option.

Binomial test: real data

Now we will load a dataset and check some null hypotheses using a binomial test.

```r
df <- read.csv("https://raw.githubusercontent.com/LingData2019/LingData/master/data/poetry_last_in_lines
               sep = "\t",
               encoding = "UTF-8")
```

Note: Before we worked with csv-files with a comma as a column separator. Now we are loading data with a tabulation sign (`\t`) used as a separator. If you work on Windows, you should specify the encoding as well since otherwise cyrillics might be unreadable.

Data info

The dataset "The last words in verses" contains a sample of lines taken from the RNC Corpus of Russian Poetry. We took only one line per author to make our observations as independent as possible.

- `Decade` - decade of creation: 1820s, 1920s
- `RhymedNwords` - the number of words in the rhyming position
- `RhymedNsyl` - the number of syllables in the rhyming position
- `UPoS` - part of speech of the last word
- `LineText` - a sampled verse
- `Author` - author of the text

Question. Suggest your hypotheses about the proportion of nouns among the words at the end of a verse.

Now let's look at frequencies:

```
table(df$UPoS)
```

```
##
##  ADJ  ADP  ADV  DET INTJ NOUN  NUM PART PRON VERB    X
##   52    1   16    6    1  222    3    2    6   54    1
```

```
table(df$UPoS)/sum(table(df$UPoS))
```

```
##
##         ADJ         ADP         ADV         DET        INTJ        NOUN
## 0.142857143 0.002747253 0.043956044 0.016483516 0.002747253 0.609890110
##         NUM        PART        PRON        VERB           X
## 0.008241758 0.005494505 0.016483516 0.148351648 0.002747253
```

The frequency of `NOUN` is 0.6 approximately. Is it enough to make conclusions about our null hypothesis? Of course, no, we could get this by chance (while choosing which verses to include in the data set), so we proceed to formal tests.

First, select rows with `NOUN` as a PoS tag:

```
nouns <- df[df$UPoS=='NOUN',]
```

Second, calculate the total number of verses (our trials like tosses of a coin) and the number of verses that end with nouns (our successes like heads):

```
total <- nrow(df) # total number of trials
nnouns <- nrow(nouns) # number of successes (success = NOUN)
```

Now we are ready to test the hypotheses you suggested!

Hypothesis: the proportion of nouns at the end of verses is 0.6.

$$H_0 : p = 0.6$$

```
# H0: p = 0.6
binom.test(nnouns, total, p = 0.6)
```

```
##
##  Exact binomial test
##
## data:  nnouns and total
## number of successes = 222, number of trials = 364, p-value =
## 0.7085
## alternative hypothesis: true probability of success is not equal to 0.6
## 95 percent confidence interval:
##  0.5576786 0.6602988
## sample estimates:
## probability of success
##              0.6098901
```

```
# not reject
```

At the 5% significance level we should not reject our null hypothesis on the data given. The true proportion of verses that ends with nouns could be 0.6.

Hypothesis: the proportion of nouns at the end of verses is 0.4.

$$H_0 : p = 0.4$$

```
# H0: p = 0.4
binom.test(nnouns, total, 0.4)
```

```
##
##  Exact binomial test
##
## data:  nnouns and total
## number of successes = 222, number of trials = 364, p-value =
## 9.059e-16
## alternative hypothesis: true probability of success is not equal to 0.4
## 95 percent confidence interval:
##  0.5576786 0.6602988
## sample estimates:
## probability of success
##              0.6098901
```

```
# reject
```

At the 5% significance level we should reject our null hypothesis on the data given. The true proportion of verses that ends with nouns is not 0.4.

Now we can choose a subset of our data frame, choose only those verses that end with one-syllable words and test whether $H_0 : p = 0.6$ should be rejected on this subset.

```
# choose lines with one-syllable words at the end
one_syll <- df[df$RhymedNsyl == 1, ]

# again save two numbers for binom.test()
total_one <- nrow(one_syll)
noun_one <- nrow(one_syll[one_syll$UPoS == "NOUN", ])

# test and make your own conclusion
binom.test(noun_one, total_one, p=0.6)
```

```
##
##  Exact binomial test
##
## data:  noun_one and total_one
## number of successes = 32, number of trials = 43, p-value = 0.06142
## alternative hypothesis: true probability of success is not equal to 0.6
## 95 percent confidence interval:
##  0.5882843 0.8648140
## sample estimates:
## probability of success
##               0.744186
```

Further you can test several hypotheses on your own. For example, for every number of syllables.

```
table(df$RhymedNsyl)
```

```
##
##   1   2   3   4   5   6
##  43 139 123  49   9   1
```