χ^2-tests in R

I. Basic uses of `chisq.test()`: Pearson's χ^2-tests

```
chisq.test( x, p = <vector-of-probabilities> )

x: a numeric vector or matrix
p: a vector of probabilities of the same length of 'x'

If `x' is a matrix with at least two rows and columns, it is taken
as a two-dimensional contingency table: the entries of `x' must be
non-negative integers.
```

Goodness-of-fit test

```
x is a vector  =>  `x' is treated as a one-dimensional contingency table

Example:
x <- c(89,37,30,28,2)
p <- c(0.40,0.20,0.20,0.15,0.05)
chisq.test(x, p = p)
```

II. Examples based on real data

Goodness-of-fit test

The data comes from the word sense disambiguation task in which the senses of the noun *line* are investigated. The estimated probabilities are relative frequencies observed in the training dataset.
The null hypothesis is that in the test dataset the senses have the same distribution. We will check the hypothesis using Pearson's χ^2-test.

1. Data

	estimated	test set
SENSES	probabilities	observations
cord	9.2%	37
division	8.9%	51
formation	8.1%	52
phone	10.6%	44
product	53.5%	268
text	9.8%	48

```
> x = c(37, 51, 52, 44, 268, 48)
> p = c(9.2, 8.9, 8.1, 10.6, 53.5, 9.8)
```

2. The formula for Pearson's cumulative test statistic

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

3. Computing the statistic in R "by hands"

```
> O = x
> E = p/100 * sum(x)

# The statistic:
> sum((O-E)*(O-E)/E)
[1] 7.525384

# The critical value of chi-square with df=5 at 95%:
> qchisq(0.95, df=5)
[1] 11.0705
```

4. Conclusion

the critical value > the computed statistic $\Longrightarrow$ we *cannot* reject the hypothesis that senses are distributed as we estimated

5. The same using `chisq.test()`

```
> chisq.test(x, p=p, rescale.p=T)
Chi-squared test for given probabilities

data:  x
X-squared = 7.5324, df = 5, p-value = 0.184
```