Part 4. Categorical variables

Natalia Levshina @ 2017
University of Mainz, June 2017

Outline

1. Counts, proportions and percentages
2. Graphs: bar plots and pie charts
3. Association between two categorical variables
   - effect size
   - Chi-squared test
   - Fisher exact test

English Lexicon Project data

```
> library(Rling)
> data(ELP)
> str(ELP)
'data.frame':	880 obs. of  5 variables:
 $ Word   : Factor w/ 880 levels "abbreviation",..: 631 747 200 773 821 134 845 140 94 354 ...
 $ Length : int  7 10 10 8 6 5 5 8 8 6 ...
 $ SUBTLWF: num  0.96 4.24 0.04 1.49 1.06 3.33 0.1 0.06 0.43 5.41 ...
 $ POS    : Factor w/ 3 levels "JJ","NN","VB": 2 2 3 2 2 2 3 2 2 2 ...
 $ Mean_RT: num  791 693 960 771 882 ...
```

2 ways to tabulate a variable

```
> attach(ELP)
> summary(POS)
JJ  NN  VB
159 532 189
> table(POS)
POS
 JJ  NN  VB
159 532 189
```

Proportions and percentages

```
> prop.table(table(POS))
POS
       JJ        NN        VB
0.1806818 0.6045455 0.2147727
> prop.table(table(POS))*100
POS
      JJ       NN       VB
18.06818 60.45455 21.47727
```

Outline

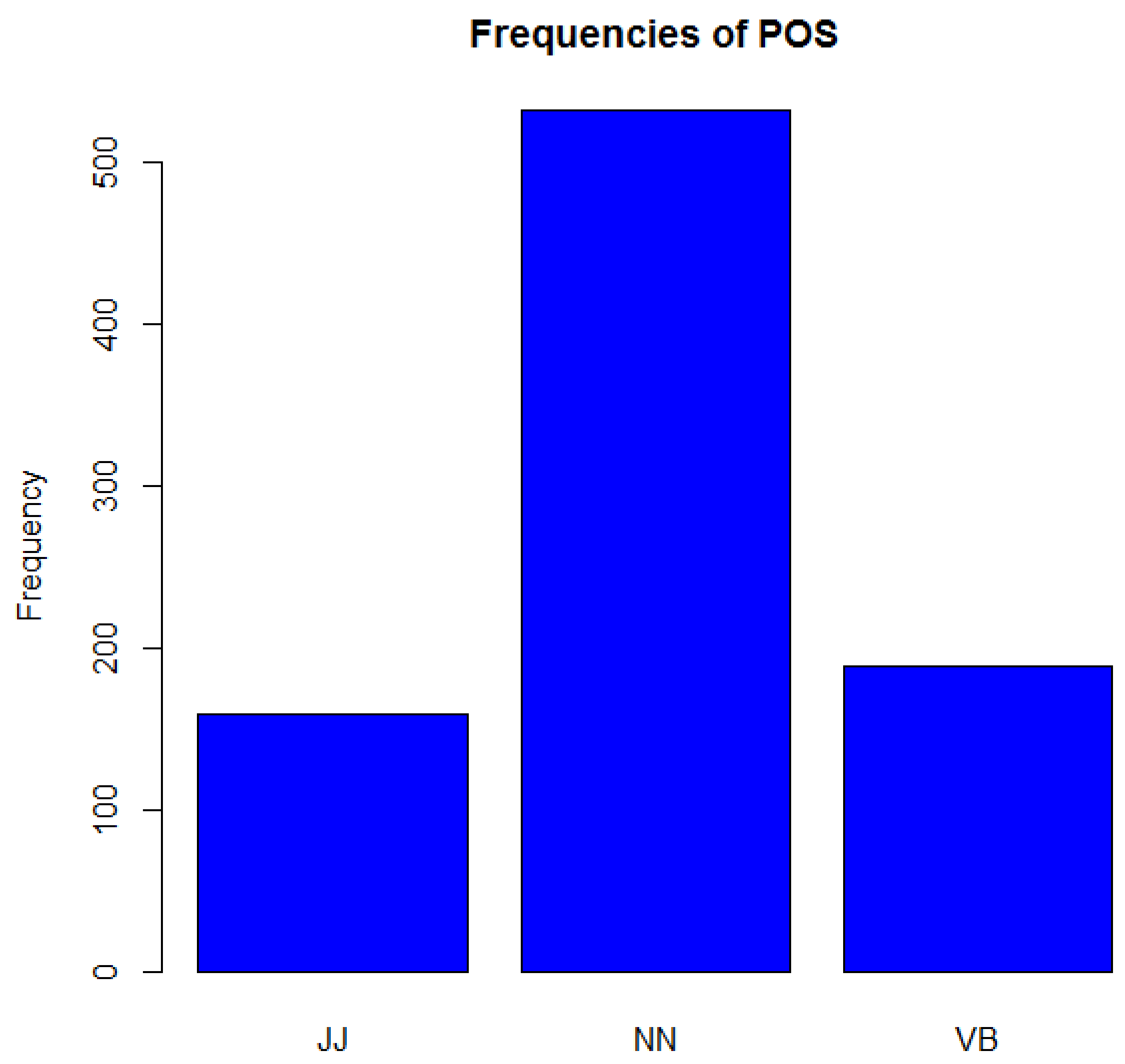
1. Counts, proportions and percentages
2. Graphs: bar plots and pie charts
3. Association between two categorical variables
    - effect size
    - Chi-squared test
    - Fisher exact test

Bar plot with counts

```
> barplot(table(POS))
```

A fancier version:

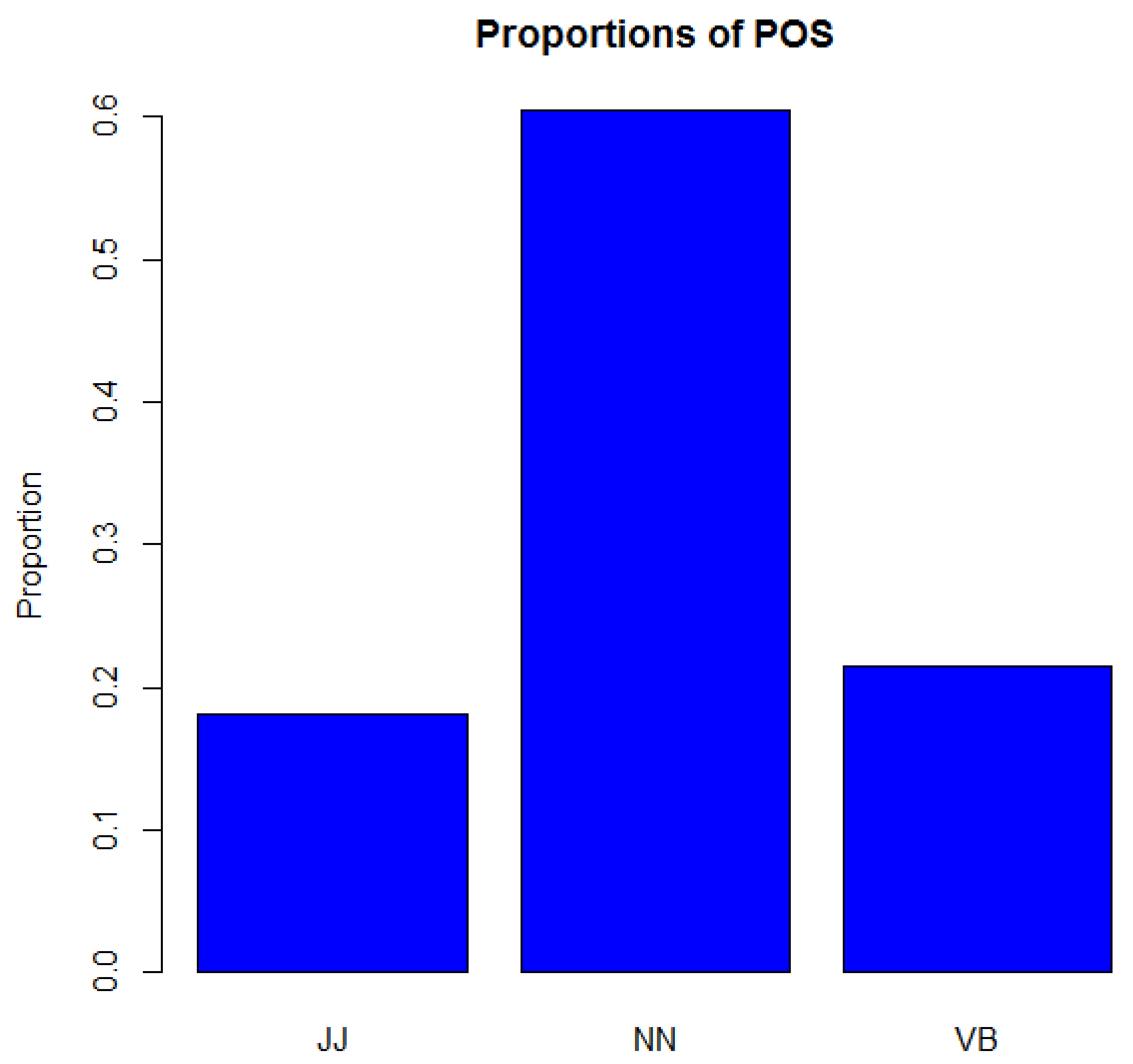
```
> barplot(table(POS), col = "blue", main =
"Frequencies of POS", ylab = "Frequency")
```



Frequencies of POS
Frequency
0
100
200
300
400
500
JJ
NN
VB

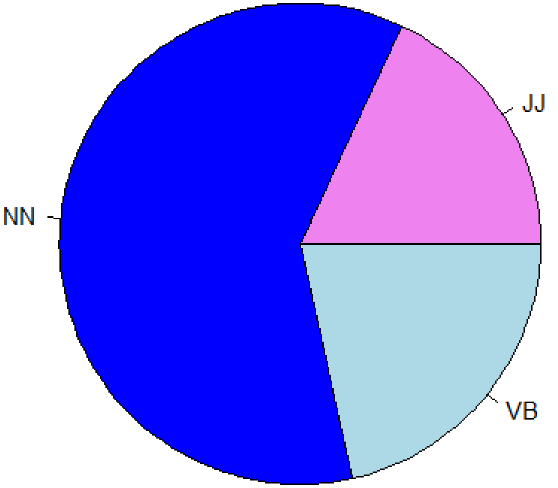
Bar plot with proportions

```
> barplot(prop.table(table(POS)), col = "blue", main =
"Proportions of POS", ylab = "Proportion")
```


Proportions of POS
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
JJ
NN
VB

Pie chart

```
> pie(table(POS), col = c("violet", "blue", "lightblue"))
```


JJ
NN
VB

Exercise: Your colours

- Create a factor with your and your fellow students' favourite colours.
- Compute the proportions of each of the colours. Which one is the most popular in the group?
- Create a pie chart with the colours corresponding to each colour category.

Detach the data

```
> detach(ELP)
```

Outline

1. Counts, proportions
2. Graphs: bar plots and pie charts
3. Association between two categorical variables
   - effect size
   - Chi-squared test
   - Fisher exact test

Nerds and geeks

```
> data(nerd)
> str(nerd)
'data.frame':	1316 obs. of  5 variables:
 $ Noun    : Factor w/ 2 levels "geek","nerd": 2 2 2 2 2 2 2 2 2 2 ...
 $ Num     : Factor w/ 2 levels "pl","sg": 1 1 1 1 1 1 1 1 1 1 ...
 $ Century : Factor w/ 2 levels "XX","XXI": 1 2 1 1 1 2 2 1 2 1 ...
 $ Register: Factor w/ 4 levels "ACAD","MAG","NEWS",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Eval    : Factor w/ 3 levels "Neg","Neutral",..: 2 2 2 2 2 2 2 2 2 2 ..
```

Noun by Century

```
> attach(nerd)
> table(Noun, Century)
      Century
Noun    XX XXI
  geek 197 473
  nerd 318 328
```

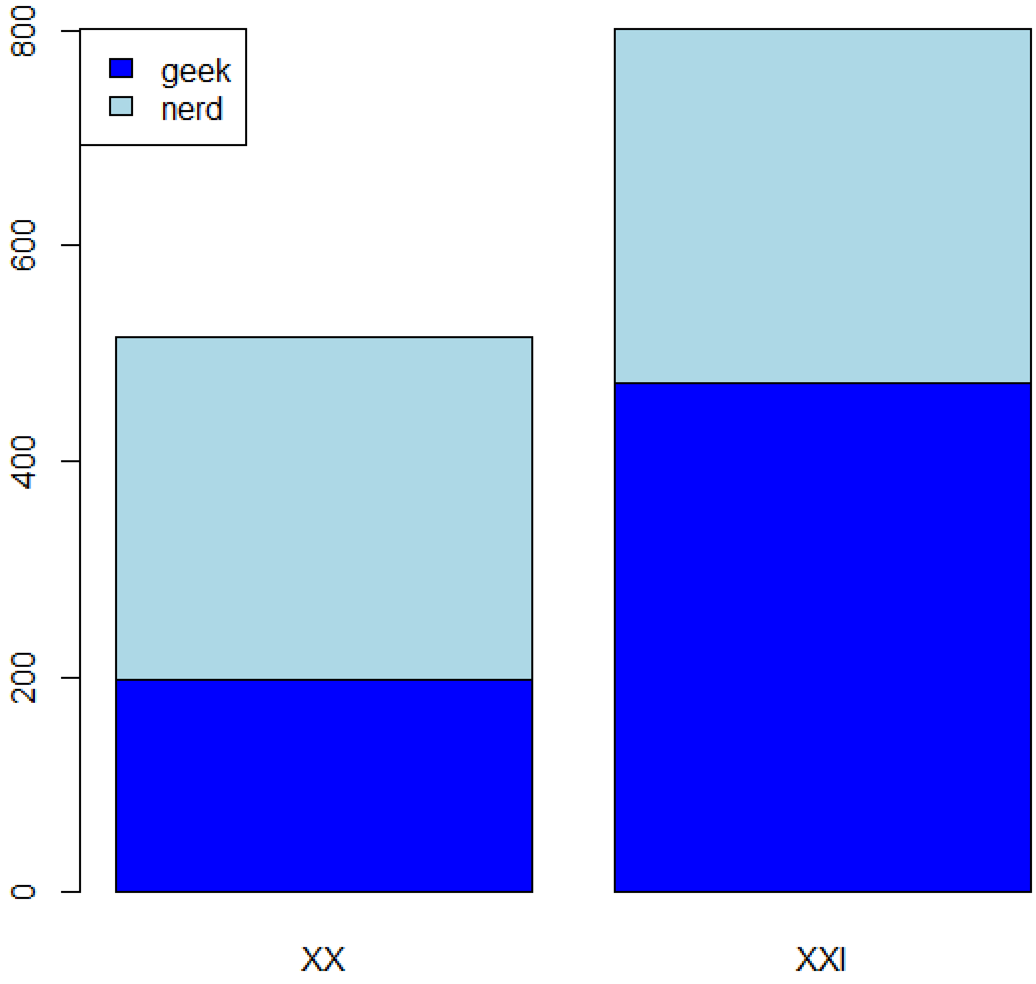
This is called a contingency table.

Proportions for two-dimensional tables

```
> prop.table(table(Noun, Century)) #all cells sum up to 1
      Century
Noun          XX       XXI
  geek 0.1496960 0.3594225
  nerd 0.2416413 0.2492401
> prop.table(table(Noun, Century), 1) #rows sum up to 1
      Century
Noun          XX       XXI
  geek 0.2940299 0.7059701
  nerd 0.4922601 0.5077399
> prop.table(table(Noun, Century), 2) #columns sum up to 1
      Century
Noun          XX       XXI
  geek 0.3825243 0.5905119
  nerd 0.6174757 0.4094881
```

Bar plots with 2 variables

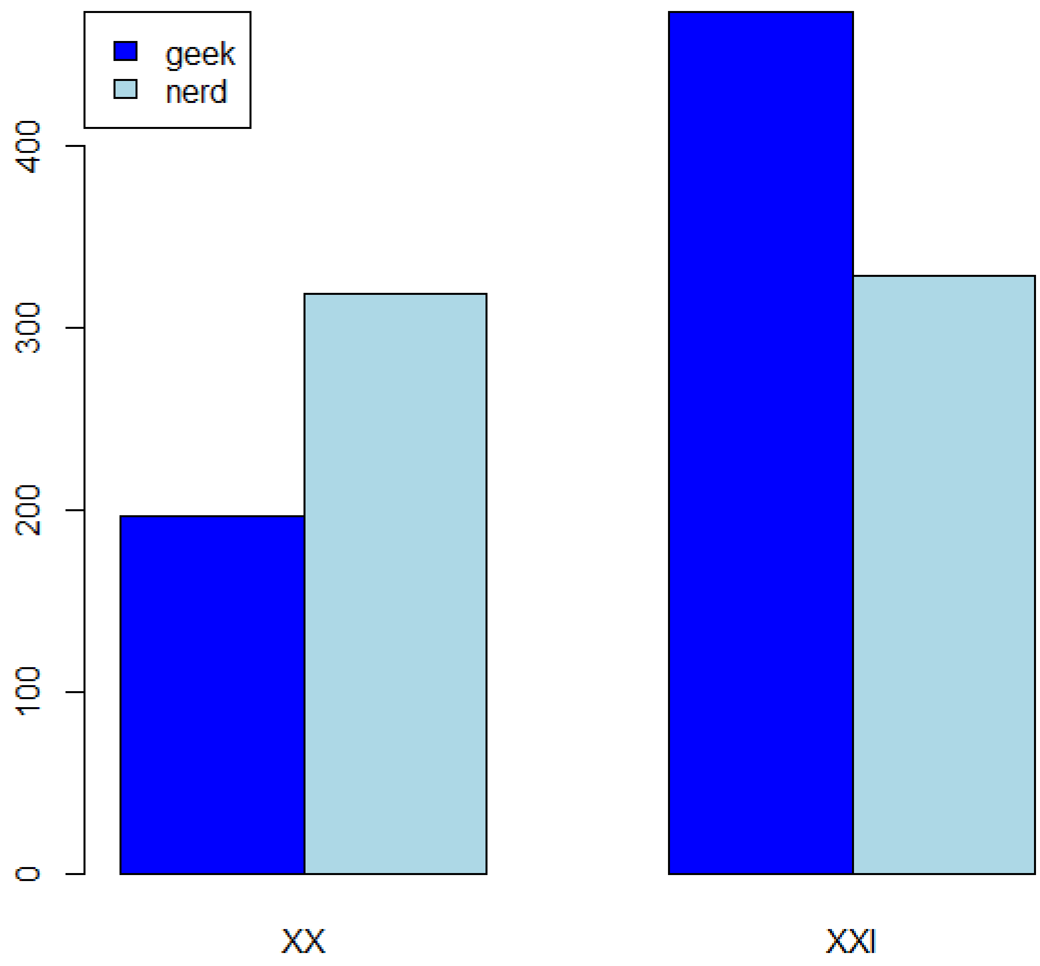
```
> barplot(table(Noun, Century), col = c("blue", "lightblue"))

> legend("topleft", legend = c("geek", "nerd"), fill = c("blue", "lightblue"))
```



800
600
400
200
0
geek
nerd
XX
XXI

Bar plots with unstacked bars

```
> barplot(table(Noun, Century), col = c("blue", "lightblue"), beside = TRUE)
> legend("topleft", legend = c("geek", "nerd"), fill = c("blue", "lightblue"))
```


geek
nerd
0
100
200
300
400
XX
XXI

Effect size for categorical data

- One can see that *nerd* occurs more often in the XX century data than in the XXI century data. For *geek*, it is the other way round.
- One speaks about an association between two categorical variables.
- How strong is that effect?
- One of the popular measures of effect size is odds ratio.

Odds and odds ratio

```
> table(Noun, Century)
     Century
Noun    XX XXI
  geek 197 473
  nerd 318 328
```

Odds geek to nerd in XX = 197/318 = 0.62

Odds geek to nerd in XXI = 473/328 = 1.44

Odds ratio (OR) = 0.62/1.44 = 0.43

How to interpret odds?

- If odds = 1, there is no difference in the probabilities of both outcomes (i.e. geek and nerd).
- If odds > 1, the probability of the first outcome (geek) is greater than the probability of the second outcome (nerd).
- If odds < 1, the probability of the first outcome (geek) is smaller than the probability of the second outcome (nerd).
- Odds ratio = 1 means no difference in the odds. No association between the variables.
- Odds ratio > 1 means that the first odds are greater than the second odds.
- Odds ratio < 1 means that the first odds are smaller than the second odds.

A cliché example

- Imagine you have 10 Belgian friends and 10 German friends. 9 of the Belgians love Belgian beers, and only 1 hates them. 9 of the German friends hate Belgian beers, and only 1 loves them.

	Belgian	**German**
Loves	9	1
Hates	1	9

- Odds of love to hate for Belgians are 9/1 = 9, and for Germans 1/9 = 0.11
- Odds ratio is 9/0.11 = 81. This is a very strong effect.

Does the order matter?

- Let’s swap the columns:

	German	**Belgian**
Loves	1	9
Hates	9	1

- OR = 0.11/9 = 0.0123
- But this is the inverse of 81!

  1/81 = 0.0123

- If an association is strong, the OR is either very close to 0 (for OR < 1), or very large (for OR > 1).

Hypothesis

- Here, the odds of *geek* against *nerd* in the XX century data are smaller than the same odds in the XXI century data.
- But is the difference statistically significant?
- Null hypothesis: there is no difference between the odds of *geek* against *nerd* in the two centuries. Or there is no association between the nouns and the centuries.
- Alternative hypothesis: there is a difference between the odds of *geek* to *nerd*. Or one can say there is an association between the nouns and the centuries.
- Note: for categorical data, it is more conventional to use non-directional hypotheses.

Outline

1. Counts, proportions
2. Graphs: bar plots and pie charts
3. Association between two categorical variables
   - effect size
   - Chi-squared test
   - Fisher exact test

Chi-squared test

```
> chisq.test(table(Noun, Century))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(Noun, Century)
X-squared = 53.429, df = 1, p-value = 2.681e-13
```

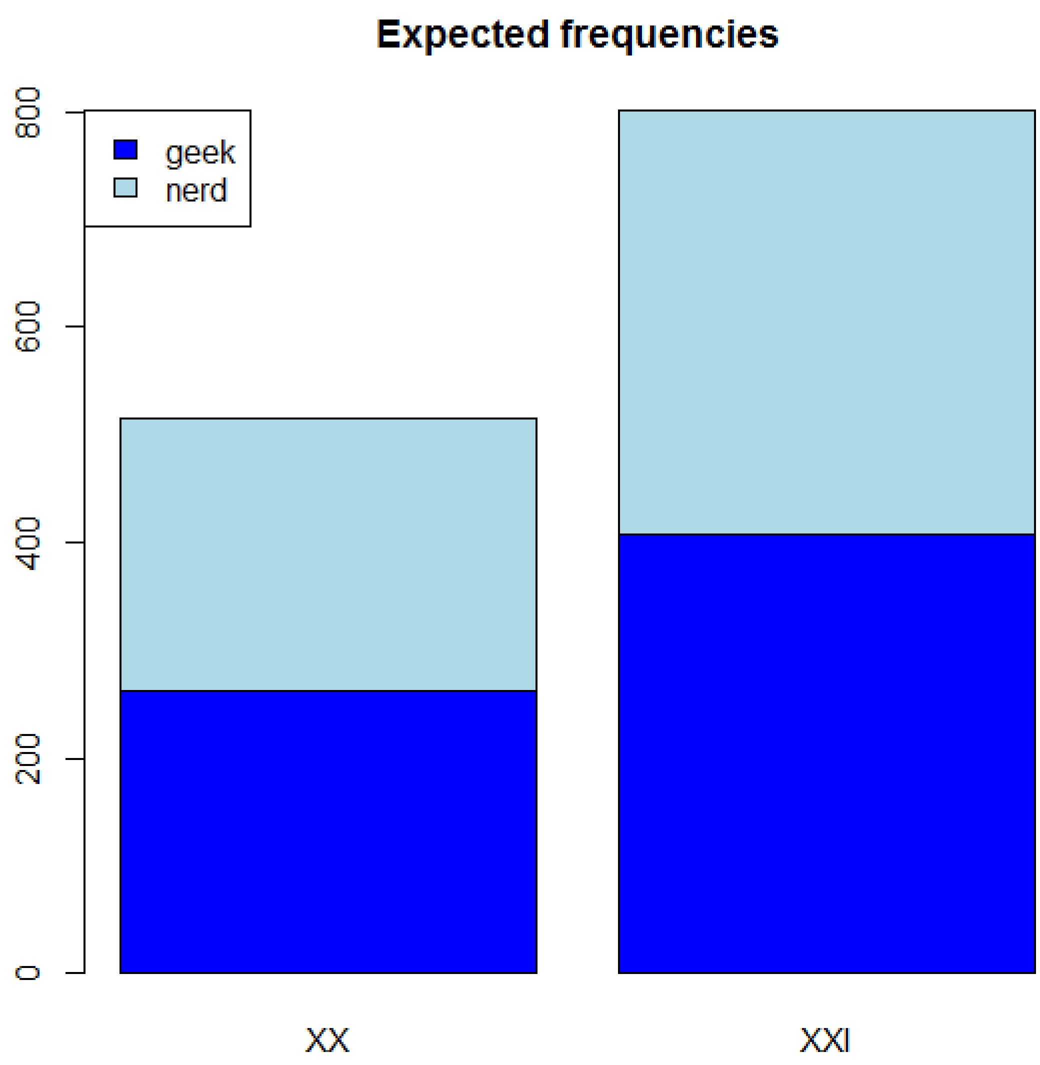
What is the Chi-squared statistic?

- A sum of squared deviations of the observed frequencies from the expected values divided by the expected values.
- The greater the deviations, the more reasons to believe that something's going on.
- The expected values are those if there is no association between the variables.

```
> chisq.test(table(Noun, Century))$expected
      Century
Noun        XX     XXI
  geek 262.196 407.804
  nerd 252.804 393.196
```

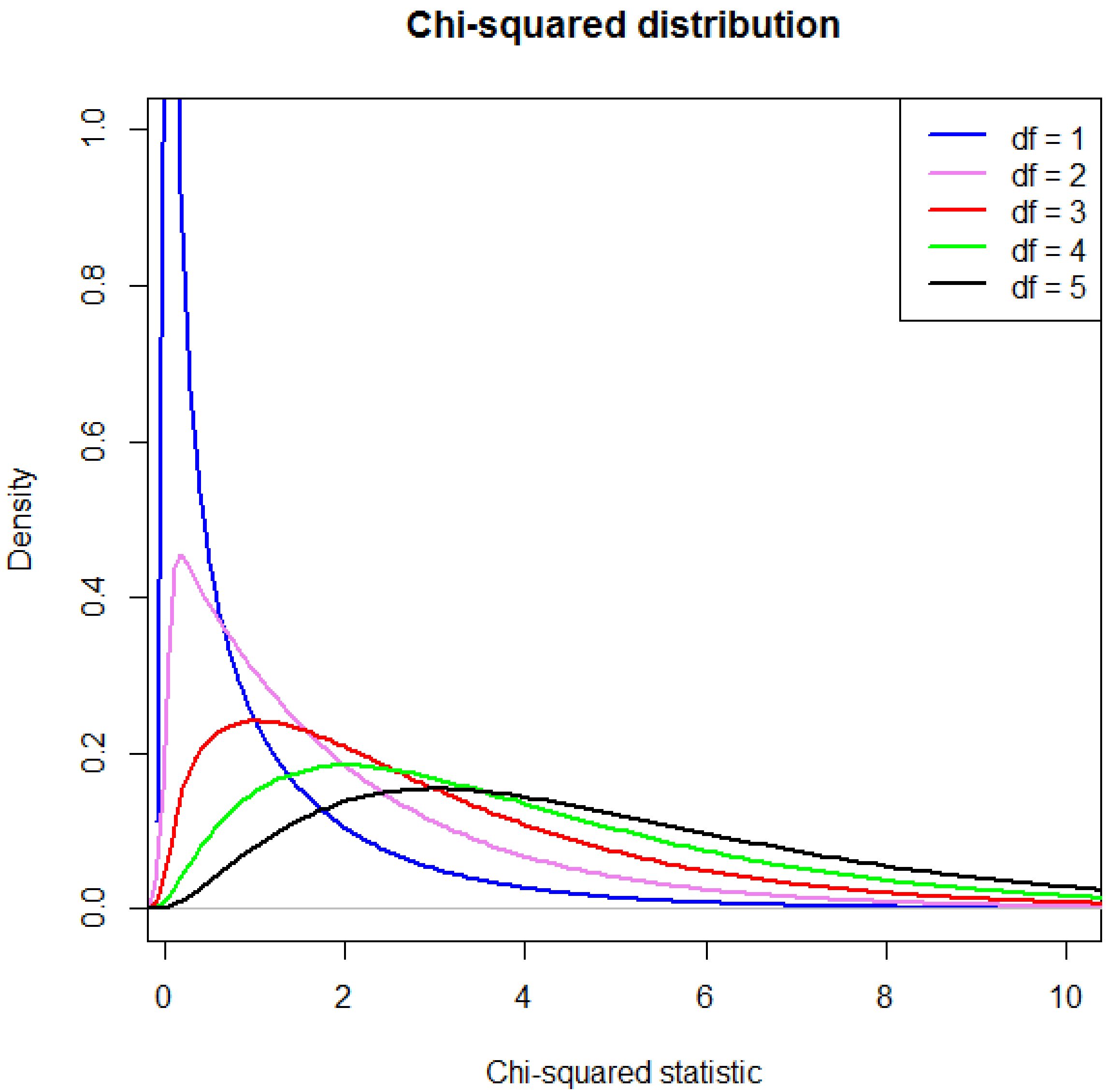
Understanding expected frequencies

```
> barplot(chisq.test(table(Noun, Century))$expected, col = c("blue", "lightblue"), main = "Expected frequencies")
> legend("topleft", legend = c("geek", "nerd"), fill = c("blue", "lightblue"))
```


Expected frequencies
800
600
400
200
0
geek
nerd
XX
XXI

What are degrees of freedom?

- Degrees of freedom are necessary for computing the *p*-value.
- For a x by y table, this is x – 1 multiplied by y – 1
- If it's a 2 x 2 table, then (2 – 1) x (2 – 1) = 1
- If it's a 3 x 3 table, then (3 – 1) x (3 – 1) = 4
- They show how many cells in the table we can change without changing the row and column sums. In our case, we can only change one cell in the table (try it out!).



Chi-squared distribution
Density
Chi-squared statistic
df = 1
df = 2
df = 3
df = 4
df = 5
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

Interpretation of the results

- A *p*-value for continuous statistics (like the Chi-squared) is computed as the proportion of the area with the given and more extreme values under the curve that corresponds to the degrees of freedom. All area = 1.
- Obviously, this area for Chi-squared ≥ 53 is tiny.
- More exactly, p = 2.681e-13, i.e. 0.0000000000002681.
- It’s highly unlikely to find this result by chance!
- We can safely reject the null hypothesis of no association.

Exercise

- Do *nerd* and *geek* differ with regard to their positive or negative evaluation? In other words, is there an association between the variables *Noun* and *Eval*?
  - Make a bar plot.
  - What is the expected frequency of nerd with negative evaluation? Is it larger or smaller than the observed frequency?
  - What is the expected frequency of geek with positive evaluation? Is it larger or smaller than the observed frequency?
  - How many degrees of freedom does the table have?
  - Perform the Chi-squared test and interpret it. Can you reject the null hypothesis of no association?

Outline

1. Counts, proportions
2. Graphs: bar plots and pie charts
3. Association between two categorical variables
   - effect size
   - Chi-squared test
   - Fisher exact test

Fisher exact test

- FET should be used in those situations when one of the expected frequencies is less than 5.
- The Chi-squared test becomes unreliable, and you get a warning.

The Fisher exact test: The story

- A biologist, Ph.D. B. Muriel Bristol-Roach, claimed that tea tasted better if the milk is added before the tea. She said she could tell the difference.
- To test this, Fisher devised an experiment:
- He gave her 8 cups of tea. In 4, the milk was added before, and in the other 4, after.
- Can she tell which cups is which, knowing that 4 + 4?

The tea challenge

- She got all eight correct!

	Said “tea first”	**Said “milk first”**
Tea first	4	0
Milk first	0	4

- Is it due to chance, or can Dr. Bristol really tell the difference?

Data set *tea*

```
> tea <- rbind(TeaFirst = c(4, 0), MilkFirst = c(0, 4))
> colnames(tea) <- c("Said_Tea", "Said_Milk")
> tea
```

	Said_Tea	Said_Milk
TeaFirst	4	0
MilkFirst	0	4

The Chi-squared test is not reliable

```
> chisq.test(tea)

	Pearson's Chi-squared test with Yates' continuity
correction

data:  tea
X-squared = 4.5, df = 1, p-value = 0.03389

Warning message:
In chisq.test(tea) :
  Chi-squared approximation may be incorrect
```

Some combinatorics

Suppose she had 0 correct guesses:

	Cup 1	**Cup 2**	**Cup 3**	**Cup 4**	**Cup 5**	**Cup 6**	**Cup 7**	**Cup 8**
Actual	T	T	T	T	M	M	M	M
Response	M	M	M	M	T	T	T	T

There is only one 1 possible combination.

Some combinatorics

If she had 1 correct guess where tea was first, there are 16 possible combinations, e.g.

	Cup 1	**Cup 2**	**Cup 3**	**Cup 4**	**Cup 5**	**Cup 6**	**Cup 7**	**Cup 8**
Actual	T	T	T	T	M	M	M	M
Response	T	M	M	M	M	T	T	T

	Cup 1	**Cup 2**	**Cup 3**	**Cup 4**	**Cup 5**	**Cup 6**	**Cup 7**	**Cup 8**
Actual	T	T	T	T	M	M	M	M
Response	M	T	M	M	M	T	T	T

And so on…

Some combinatorics

If she had 2 correct guesses where tea was first, there are 36 possible combinations, e.g.

	Cup 1	**Cup 2**	**Cup 3**	**Cup 4**	**Cup 5**	**Cup 6**	**Cup 7**	**Cup 8**
Actual	T	T	T	T	M	M	M	M
Response	T	T	M	M	M	M	T	T

	Cup 1	**Cup 2**	**Cup 3**	**Cup 4**	**Cup 5**	**Cup 6**	**Cup 7**	**Cup 8**
Actual	T	T	T	T	M	M	M	M
Response	M	T	T	M	M	M	T	T

And so on…

Some combinatorics

- If she guesses 3 cups where the tea was added first, there are again 16 combinations (if you don’t believe, try it out!).
- A terribly difficult question: How many combinations are there for guessing correctly?

Permutations

In total, we have

1 + 16 + 36 + 16 + 1 = 70 possible combinations, or permutations.

There is only one combination when all cups are guessed correctly.

$$P = 1/70 \approx 0.0143$$

This is the exact *p*-value for a one-tailed test.

FET in R (one-tailed)

```
> fisher.test(tea, alternative = "greater")

        Fisher's Exact Test for Count Data

data:  tea
p-value = 0.01429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 2.003768      Inf
sample estimates:
odds ratio
       Inf
```

FET in R (two-tailed)

```
> fisher.test(tea)

        Fisher's Exact Test for Count Data

data:  tea
p-value = 0.02857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.339059      Inf
sample estimates:
odds ratio
       Inf
```

Take-home messages

1. Use visualization tools.
2. For 2 by 2 tables, report the odds ratio (effect size measure). Unfortunately, it doesn’t make sense for larger tables.
3. If one of your expected frequencies is smaller than 5, use the Fisher Exact Test.
4. There’s no harm in using FET in other situations if you have not too many counts in total.