Biostatistics
Week 6

- **Risk, Odds, Risk Ratio (RR), Odds Ratio (OR)**

- **Observational studies**
  - **cohort study**
  - **case control study**
  - **cross-sectional study**

- **Controlled experiments**
  - **randomized controlled trial (RCT)**
  - **fix all potential influence factors in lab expriments**

- **Tests for change of distribution of a categorial variable**
  - **Chi-square test**
  - **Fischer-exact test**

Risk, Risk-Ratio: the classical definition

Example: We look at the prediction-power of X: “heavy passive smoker” results for the risk for heart disease (y=1) by analyzing a cross-table based on a representative sample with n = 52 subjects (data are made up).

$$risk: \; P(y=1|x)$$

$$p(y=1|x=1) = \frac{11}{11+4} = \frac{11}{15} = 73\%$$

$$p(y=1|x=0) = \frac{9}{9+28} = \frac{9}{37} = 24\%$$

response	predictor $x=0$	predictor $x=1$	
$y=0$	28	4	32
$y=1$	9	11	20
	37	15	52

$$\text{risk-ratio}: \; RR = \frac{P(y=1|x=1)}{P(y=1|x=0)}$$

$$RR = \frac{0.73}{0.24} = 3$$

The relative risk of 3 tells us, that the risk for heart-disease in subjects who are heavy passive smokers is 3-times higher than in subjects who do not active or passive smoke.

Odds and Odds-Ratio: another measure for risk and relative risk

The odds («Wettverhältnis»)
is the ratio of the probability for heart disease to the probability of no-heart disease given a certain value of the predictor.

$$odds(x) = \frac{p(y=1|x)}{p(y=0|x)} = \frac{p(y=1|x)}{1-p(y=1|x)} = \frac{p}{1-p}$$

response	predictor $x=0$	predictor $x=1$	
$y=0$	28	4	32
$y=1$	9	11	20
	37	15	52

$$odds(x=1) = \frac{0.73}{1-0.73} = 2.7 \qquad odds(x=0) = \frac{0.24}{1-0.24} = 0.3$$

$$odds-ratio:\ OR(x) = \frac{odds(x=1)}{odds(x=0)} = \frac{2.7}{0.3} = 8.6$$

The odds-ratio of 8.6 tells that the odds (risk-measure) for heart disease is 8.6 times higher in the case of subject who are heavy passive smokers than in subjects who do not active or passive smoke.

Reminder: Study in Caerphilly (Wales), 1979-2003

Prospective observational study with 914 healthy men, between 45 and 95 years who were followed over 10 years.

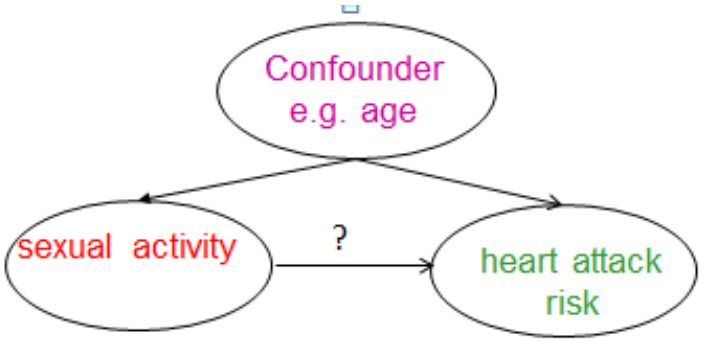
Summary:

group	# men	# sexual active men	# sexual inactive men
all men	914	231	197
men suffering heart attack	105	(8%) 19	(17%) 33

$$RR = \frac{P(y=1 \mid x=1)}{P(y=1 \mid x=0)} = \frac{0.08}{0.17} = 0.47$$

-> The risk for heart attack is in the group of sexual active men roughly 50% as in the group of sexual inactive men.

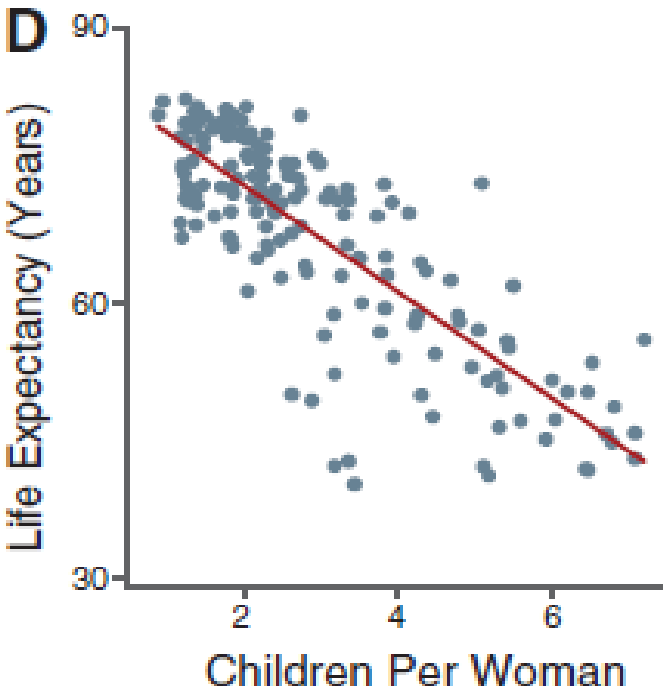
In observational studies observed association might always be caused by confounders.



A controlled experiment would allow to infer about causality.

Ecological bias

WHO has collected 357 variables (e.g. life expectancy and #children per woman) for 202 countries.



This negative association does probably not imply that a woman tends to die earlier if she gets more children.

An association on aggregated data does in general not imply (the same) association on the individual level!

Modern epidemiology is based on observational studies on individual data

- Analysis of individual data
- Ecological bias can be avoided
- Studies may be expensive and difficult to conduct

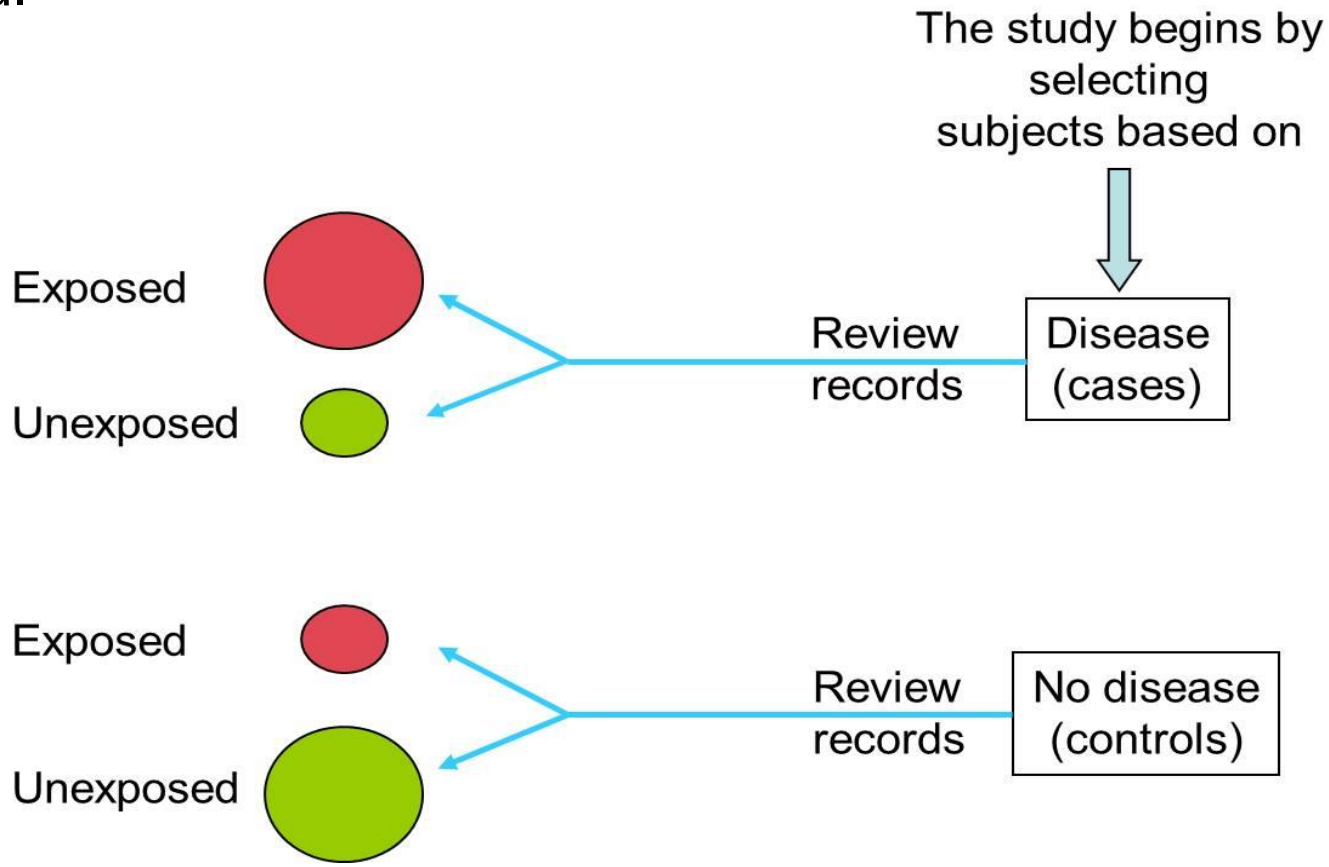
- Modern epidemiology starts around 1950 with the first cohort and case-control studies

- Motivation: dramatic increase of chronic disease mortality (cancer, cardio-vascular diseases)

- Recently there has been increasing interest in infectious disease epidemiology (z.B. AIDS, SARS, H1N1)

Case-Control Study: A first glimpse

Case-Control Study:
Retrospective study where a group of diseased persons and a control group of healthy persons are compared with respect to the exposure to a certain risk factor in the past. A disease based sampling is used, since the groups are chosen such that a certain number of diseased and healthy individuals are compared.



Picture source: http://phprimer.afmc.ca/

Case-Control study:
Which probabilities can we estimate from the resulting cross table?

	D	$\bar{D}$	
E	$n_{11} = a$	$n_{12} = \mathrm{b}$	$n_E = n_{1\bullet}$
$\bar{E}$	$n_{21} = c$	$n_{22} = d$	$n_{\bar{E}} = n_{2\bullet}$
	$n_D = n_{\bullet 1}$	$n_{\bar{D}} = n_{\bullet 2}$	n

Case-Control Study uses disease based sampling ~>

$n_D = \mathrm{n}_{\bullet 1}$ and $n_{\bar{D}} = n_{\bullet 2}$ is predefined and fix

With this kind of data, it is not possible to estimate marginal or joint population probabilities $P(D),\ P(\bar{D}),\ P(\mathrm{E}),\ P(\bar{\mathrm{E}}),\ P(E \cap D),\ldots$

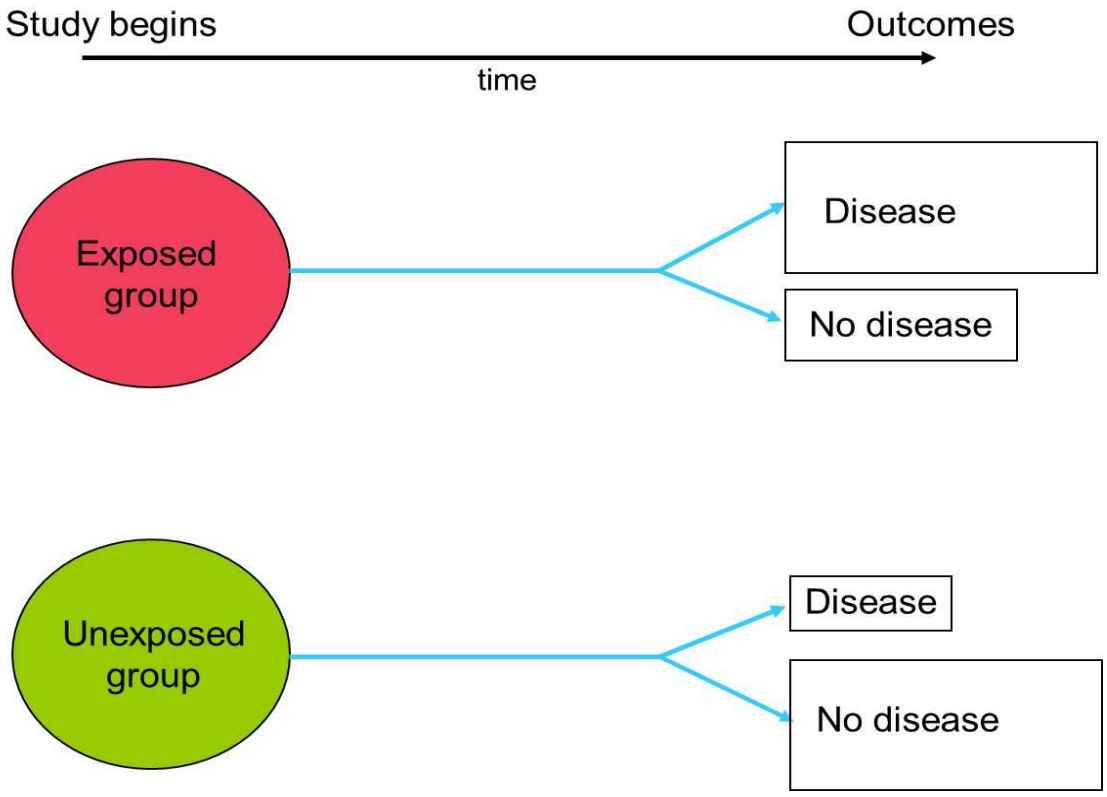
But we can estimate probabilities **conditioned on disease status**

$$\hat{P}(E \mid D) = \frac{a}{a+c} \quad , \quad \hat{P}(E \mid \bar{D}) = \frac{b}{b+d} \quad \ldots \qquad OR_E = \frac{\dfrac{P(\mathrm{E} \mid \mathrm{D})}{P(\bar{\mathrm{E}} \mid \mathrm{D})}}{\dfrac{P(\mathrm{E} \mid \bar{\mathrm{D}})}{P(\bar{\mathrm{E}} \mid \bar{\mathrm{D}})}} = \frac{a \cdot d}{b \cdot c}$$

Cohort study: A first glimpse

Cohort Study:

A exposure based sampling is used, since the cohort is chosen such that a certain number of exposed and unexposed individuals are contained in the cohort. The design is prospective since after sampling the cohort is observed over time and the future incidences of disease are compared between exposed and unexposed persons.



Picture source: http://phprimer.afmc.ca/

Cohort study:
Which probabilities can we estimate from the resulting cross table?

	D	$\bar{D}$	
E	$n_{11} = a$	$n_{12} = \mathrm{b}$	$n_E = n_{1\bullet}$
$\bar{E}$	$n_{21} = c$	$n_{22} = d$	$n_{\bar{E}} = n_{2\bullet}$
	$n_D = n_{\bullet 1}$	$n_{\bar{D}} = n_{\bullet 2}$	n

Cohort Study uses exposure based sampling ~>

$n_E = \mathrm{n}_{1\bullet}$ and $n_{\bar{E}} = n_{2\bullet}$ is predefined and fix

With this kind of data, it is not possible to estimate marginal or joint population probabilities $P(\mathrm{E})$, $P(\bar{\mathrm{E}})$, $P(D)$, $P(\bar{D})$, $P(E \cap D)$,...

But we can estimate probabilities **conditioned on exposure**

$$\hat{P}(\mathrm{D} \mid \mathrm{E}) = \frac{a}{a+b} \quad ,$$

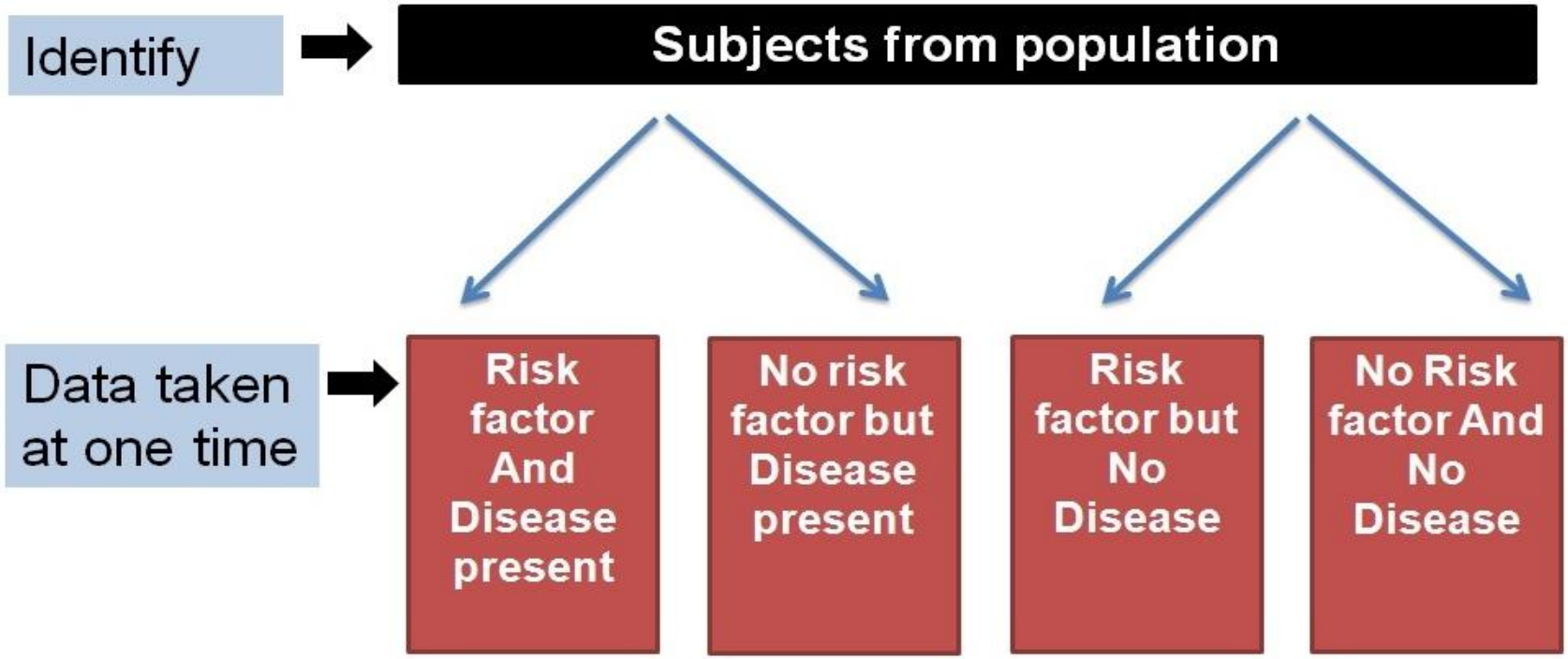
$$\hat{P}(\mathrm{D} \mid \bar{\mathrm{E}}) = \frac{c}{c+d} \quad ...$$

$$RR = \frac{P(\mathrm{D} \mid \mathrm{E})}{P(\mathrm{D} \mid \bar{\mathrm{E}})}$$

$$OR = \frac{\frac{P(\mathrm{D} \mid \mathrm{E})}{P(\bar{\mathrm{D}} \mid \mathrm{E})}}{\frac{P(\mathrm{D} \mid \bar{\mathrm{E}})}{P(\bar{\mathrm{D}} \mid \bar{\mathrm{E}})}} = \frac{a \cdot d}{b \cdot c}$$

Cross-sectional study: A first glimpse

Population based or cross-sectional study:
The sampling is done population based. A random sample of the population is taken and the current exposure and disease status of the individuals are recorded.



Picture source: http://blog.manuscriptedit.com/

Cross-sectional study
Which probabilities can we estimate from a cross table?

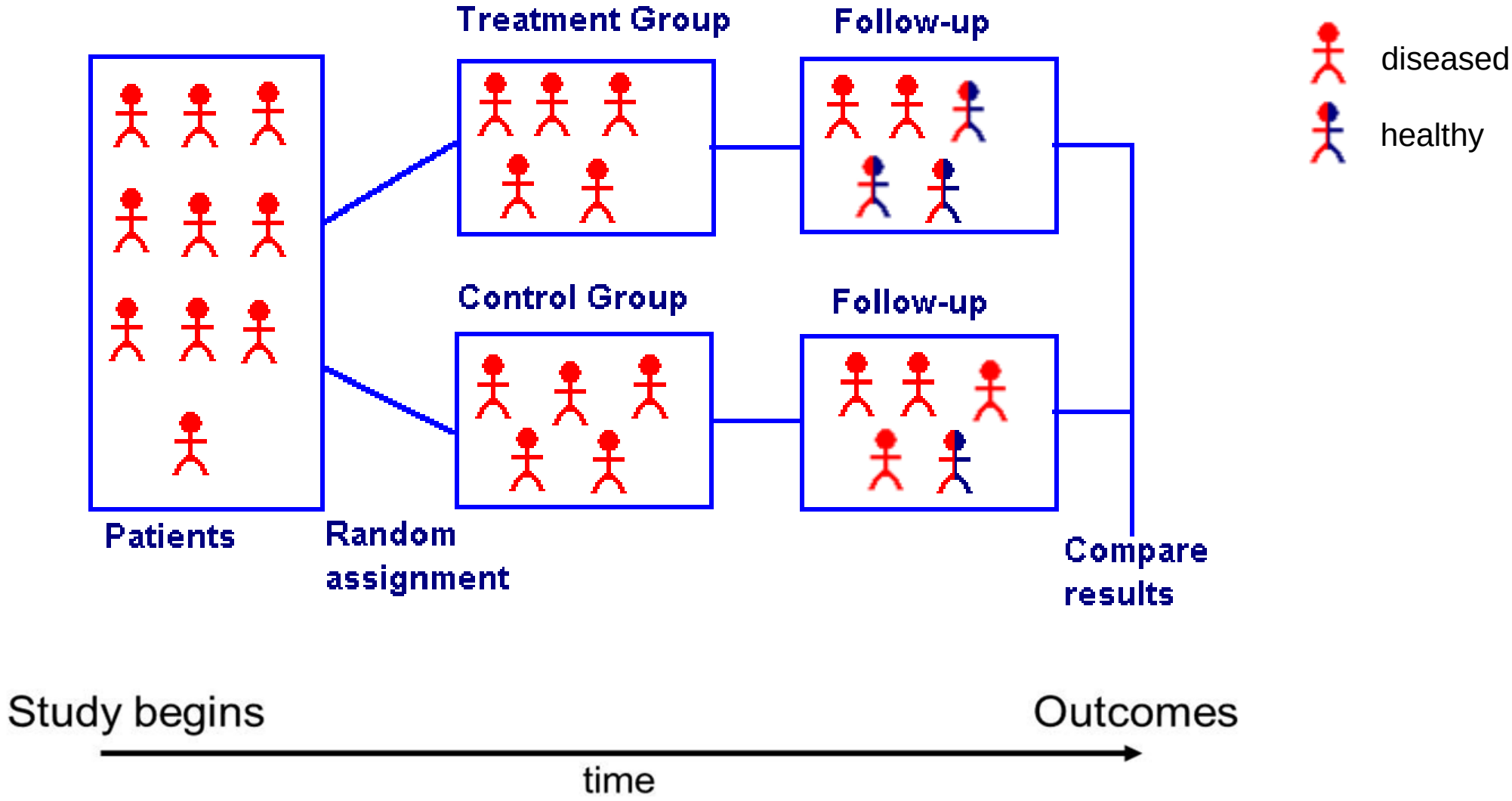
	D	$\bar{D}$	
E	$n_{11} = a$	$n_{12} = \mathrm{b}$	$n_E = n_{1\bullet}$
$\bar{E}$	$n_{21} = c$	$n_{22} = d$	$n_{\bar{E}} = n_{2\bullet}$
	$n_D = n_{\bullet 1}$	$n_{\bar{D}} = n_{\bullet 2}$	$\boldsymbol{n}$

Population based sampling ~> only n is fixed

With this kind of data, we can estimate all marginal, joint or conditional population probabilities $P(\mathrm{E}),\ P(\bar{\mathrm{E}}),\ P(D),\ P(\bar{D}),\ P(E \cap D),\ P(E \mid D), \ldots$

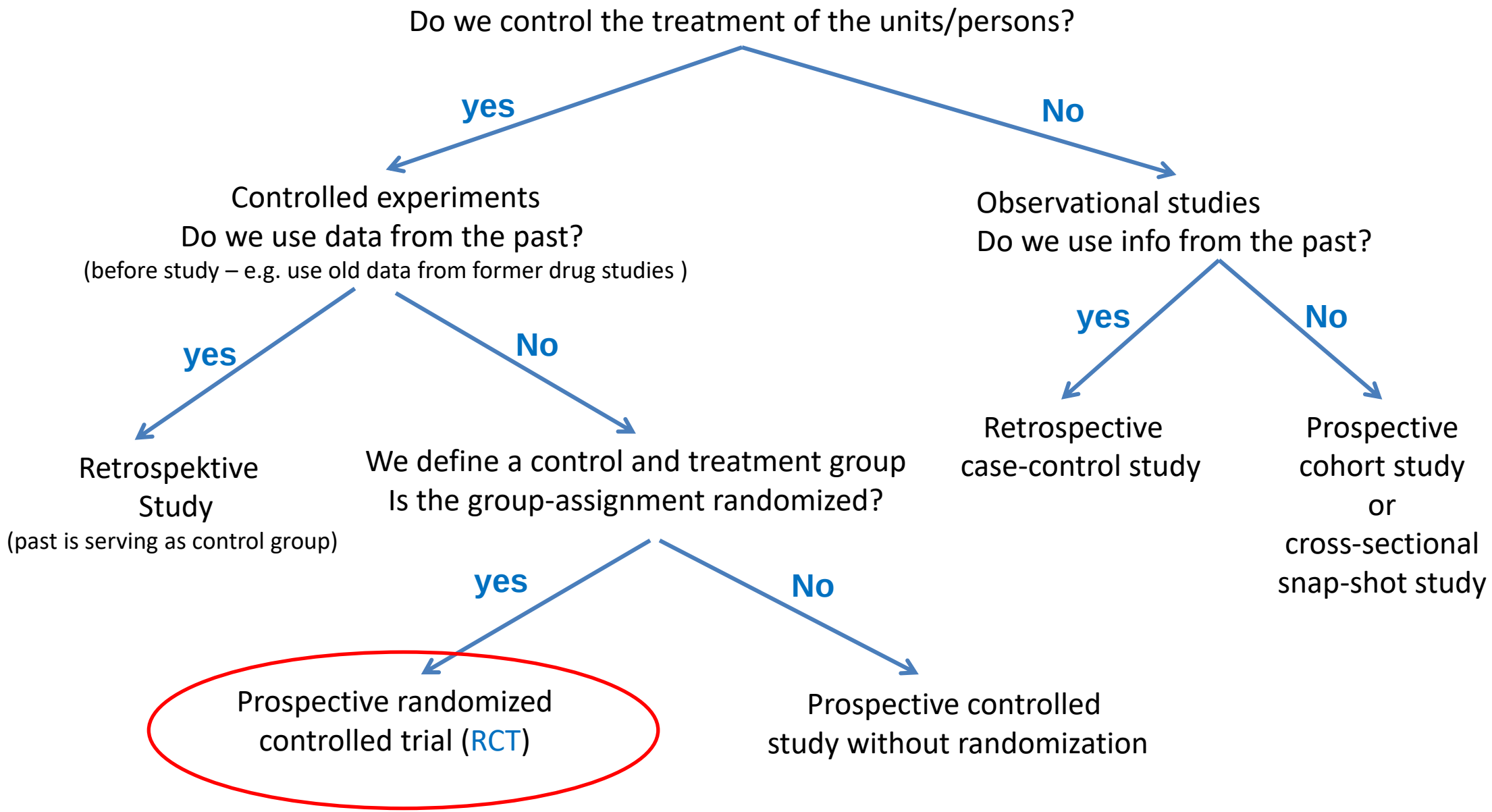
$$\hat{P}(E \cap D) = \frac{a}{n} \qquad \hat{P}(D) = \frac{a+c}{n} \qquad \hat{P}(E \mid D) = \frac{a}{a+c} \quad \ldots \qquad OR = \frac{a \cdot d}{b \cdot c} \quad \ldots$$

A clinical trial is a controlled experiments allowing for causal inference: Randomized controlled trial



Source: http://nutrevolve.blogspot.ch/2015/07/considering-evidence-based-nutrition-rct.html

Study types with human subjects



Remark to RCT: In blinded RCT the study subjects do not know if they get the drug or the placebo.

Measure of Disease-Exposure Association: Relative Risk

The *Relative Risk* or *Risk Ratio (RR)* for an outcome *D* associated with a binary risk factor *E*, is defined as follows:

Reminder: risk π=P(D)

$$RR = \frac{P(\mathrm{D} \mid \mathrm{E})}{P(\mathrm{D} \mid \bar{\mathrm{E}})}$$

Can **not** be estimated from a case-control study

$$RR = \begin{cases} > 1 \text{ . if the exposure is associated with increased disease risk} \\ = 1 \text{ , if disease risk is not associated with exposure} \\ < 1 \text{ . if the exposure is associated with decreased disease risk} \end{cases}$$

The *RR is has a lower and upper limit*:

$$0 \le P(D \mid \overset{(-)}{E}) \le 1 \quad \Rightarrow \quad 0 \ \le \ RR \ \le \ \frac{1}{P(\mathrm{D} \mid \bar{\mathrm{E}})}$$

The *RR is* **not** *symmetric in the role of the two factors D* and *E*:

$$\frac{P(\mathrm{D} \mid \mathrm{E})}{P(\mathrm{D} \mid \bar{\mathrm{E}})} \neq \frac{P(\mathrm{E} \mid D)}{P(E \mid \bar{D})}$$

Measure of Disease-Exposure Association: Odds Ratio

The *Odds Ratio (OR)* for an outcome *D* associated with a binary risk factor *E,* is defined as follows:

Reminder: $odds = \frac{P(D)}{P(\bar{D})}$

$$OR = \frac{\frac{P(\mathrm{D}\,|\,\mathrm{E})}{P(\bar{\mathrm{D}}\,|\,\mathrm{E})}}{\frac{P(\mathrm{D}\,|\,\bar{\mathrm{E}})}{P(\bar{\mathrm{D}}\,|\,\bar{\mathrm{E}})}}$$

Can be estimated from cohort **and** case-control studies

$$OR = \begin{cases} > 1 \text{ . if the exposure is associated with increased disease risk} \\ = 1 \text{ , if disease risk is not associated with exposure} \\ < 1 \text{ . if the exposure is associated with lowerded disease risk} \end{cases}$$

The *OR has only a lower bound*: $0 \le OR \le \infty$

The *OR is symmetric in the role of the two factors D* and *E*:

$$OR_D = \frac{\frac{P(\mathrm{D}\,|\,\mathrm{E})}{P(\bar{\mathrm{D}}\,|\,\mathrm{E})}}{\frac{P(\mathrm{D}\,|\,\bar{\mathrm{E}})}{P(\bar{\mathrm{D}}\,|\,\bar{\mathrm{E}})}} = \frac{\frac{P(\mathrm{E}\,|\,\mathrm{D})}{P(\bar{\mathrm{E}}\,|\,\mathrm{D})}}{\frac{P(\mathrm{E}\,|\,\bar{\mathrm{D}})}{P(\bar{\mathrm{E}}\,|\,\bar{\mathrm{D}})}} = OR_E$$

Symmetry of roles of disease and exposure in the odds ratio

$$OR_D = \frac{\dfrac{P(D\mid E)}{P(\bar{D}\mid E)}}{\dfrac{P(D\mid \bar{E})}{P(\bar{D}\mid \bar{E})}} = \frac{\dfrac{P(D\cap E)/P(E)}{P(\bar{D}\cap E)/P(E)}}{\dfrac{P(D\cap \bar{E})/P(\bar{E})}{P(\bar{D}\cap \bar{E})/P(\bar{E})}} = \frac{\dfrac{P(D\cap E)}{P(\bar{D}\cap E)}}{\dfrac{P(D\cap \bar{E})}{P(\bar{D}\cap \bar{E})}}$$

$$\|\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\|$$

$$OR_E = \frac{\dfrac{P(E\mid D)}{P(\bar{E}\mid D)}}{\dfrac{P(E\mid \bar{D})}{P(\bar{E}\mid \bar{D})}} = \frac{\dfrac{P(D\cap E)/P(D)}{P(\bar{D}\cap E)/P(D)}}{\dfrac{P(D\cap \bar{E})/P(D)}{P(\bar{D}\cap \bar{E})/P(D)}} = \frac{\dfrac{P(D\cap E)}{P(\bar{D}\cap E)}}{\dfrac{P(D\cap \bar{E})}{P(\bar{D}\cap \bar{E})}}$$

How to test if two categorial variables are independent?

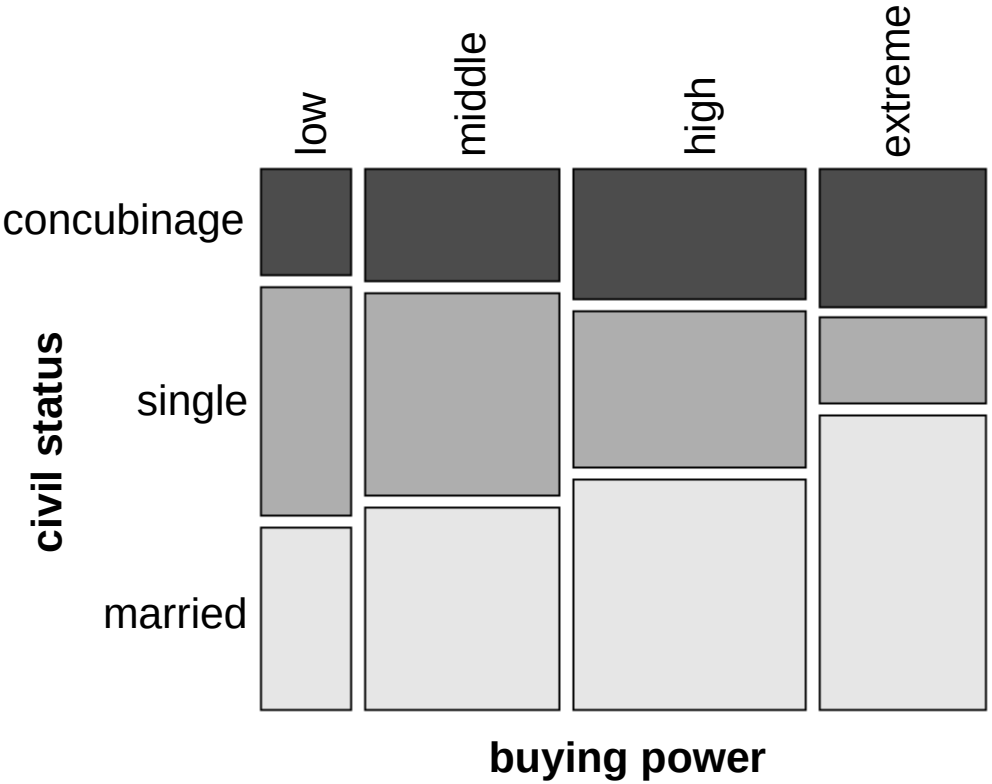
How to assess if there is a change of a categorial outcome variable when explanatory variables (treatment) change?

- **Tests for change of distribution of a categorial variable based on cross-tables**
  - **Chi-square test**
  - **Fischer-exact test**

- **Regression and classification methods (see later lectures)**

Are civil status and buying power associated?

buying power	civil status: concubinage	single	married
low	457	978	781
middle	1044	1885	1870
high	1452	1748	2550
extreme	1091	695	2330



Can you think of a confounding factor?

Independence of row and column. Independence means that knowing the value of the row variable does not change the probabilities of the column variable (and vice versa).

Or: The row percentages (or column percentages) remain constant from row to row (or column to column).

The χ^2 test for RxC crosstables

		Column Variable					
		Column 1	⋯	**Column *j***	⋯	**Column *C***	**Total**
Row Variable	**Row 1**	O_{11}	⋯	O_{1j}	⋯	O_{1C}	$n_{1\cdot}$
	⋮	⋮	⋱	⋮	⋰	⋮	⋮
	Row *i*	O_{i1}	⋯	O_{ij}	⋯	O_{iC}	$n_{i\cdot}$
	⋮	⋮	⋰	⋮	⋱	⋮	⋮
	Row *R*	O_{R1}	⋯	O_{Rj}	⋯	O_{RC}	$n_{R\cdot}$
	Total	$n_{\cdot 1}$	⋯	$n_{\cdot j}$	⋯	$n_{\cdot C}$	n

H_0: Row- and column-factors are independent

Observed value: O_{ij}

Expected value: $E_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$

$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{\text{approximativ}}{\underset{\text{under } H_0}{\sim}} \chi^2_{df=(R-1)\cdot(C-1)}$$

The approximation of the chi-square test is only valid, of number in all cells are >5.

The χ^2 distribution's shape depends on the degree of freedom k

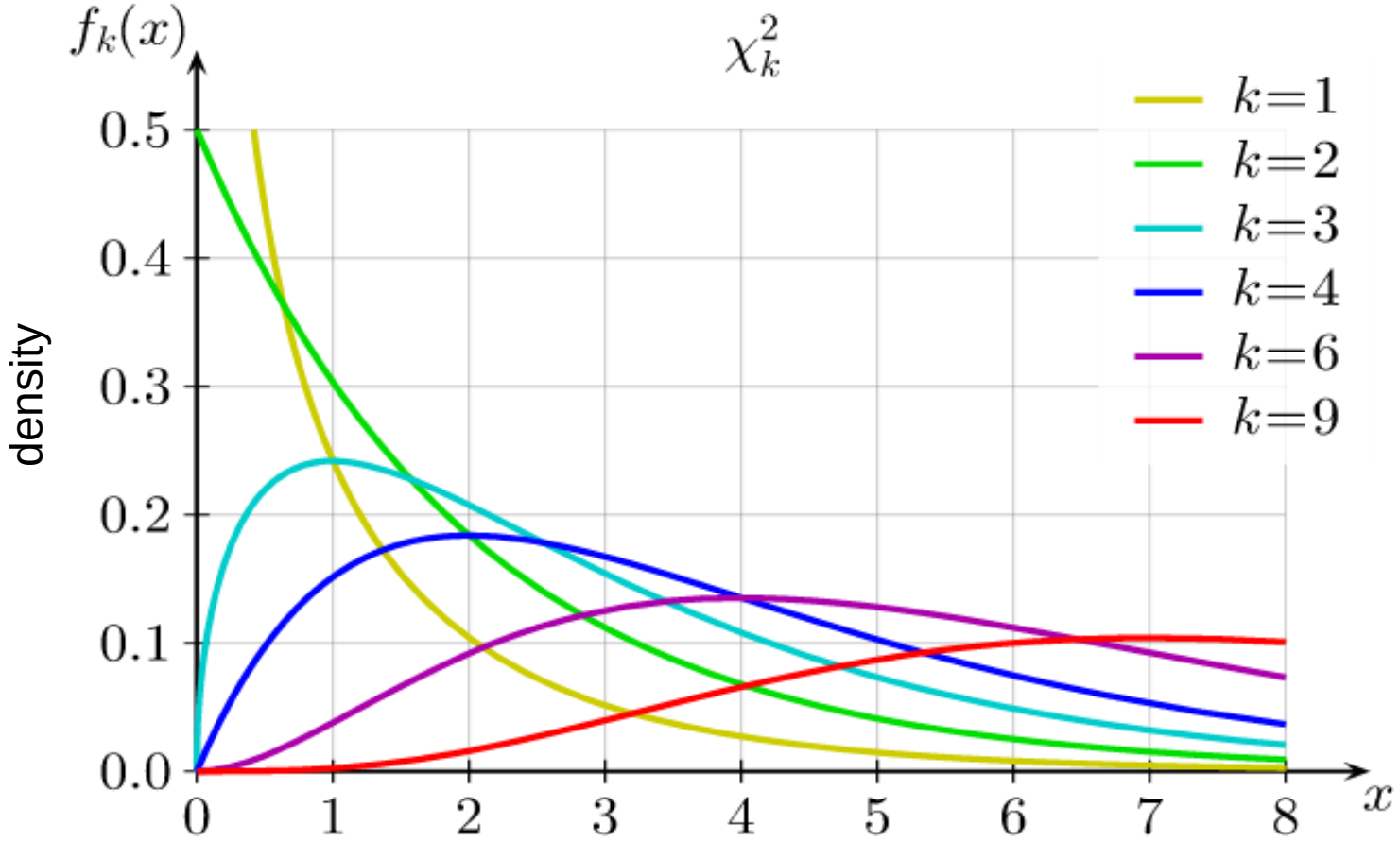


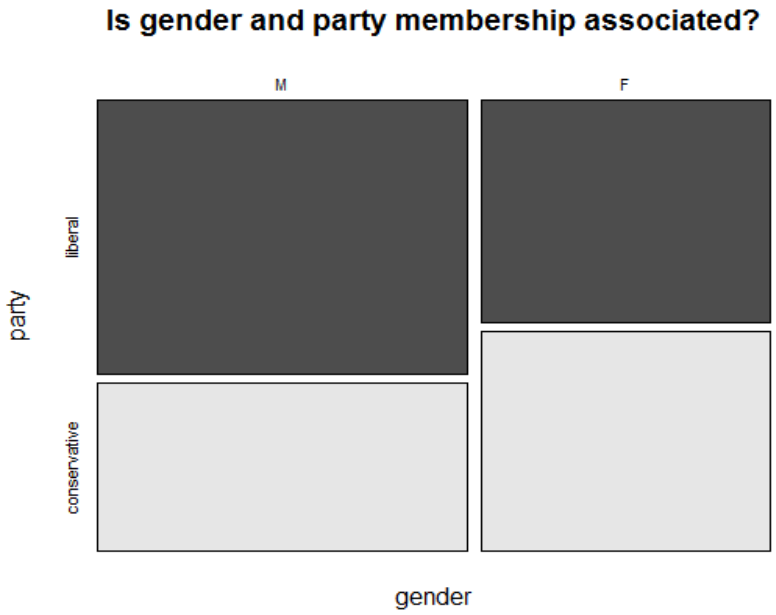
Image taken from
https://commons.wikimedia.org/w/index.php?curid=9884213

Expected value: $\mathrm{E}(X) = k$

Variance: $\mathrm{Var}(X) = 2 \cdot k$

Are gender and political orientation associated?
H_0: the two factors are independent

gender	party liberal	conservative
M	762	468
F	484	477




```
my.tab <- as.table(rbind(c(762, 468),
                         c(484, 477)))
dimnames(my.tab)=list(gender = c("M","F"),
       party = c("liberal","conservative"))
plot(my.tab,col=TRUE)
(Xsq <- chisq.test(my.tab))  # Prints test summary
Xsq$observed   # table of observed counts (same as my.tab)
Xsq$expected   # table of expected counts under the null
Xsq$residuals  # table of Pearson residuals (obs-exp)/sqrt(exp)
```

Gives insight in deviations from independence assumption

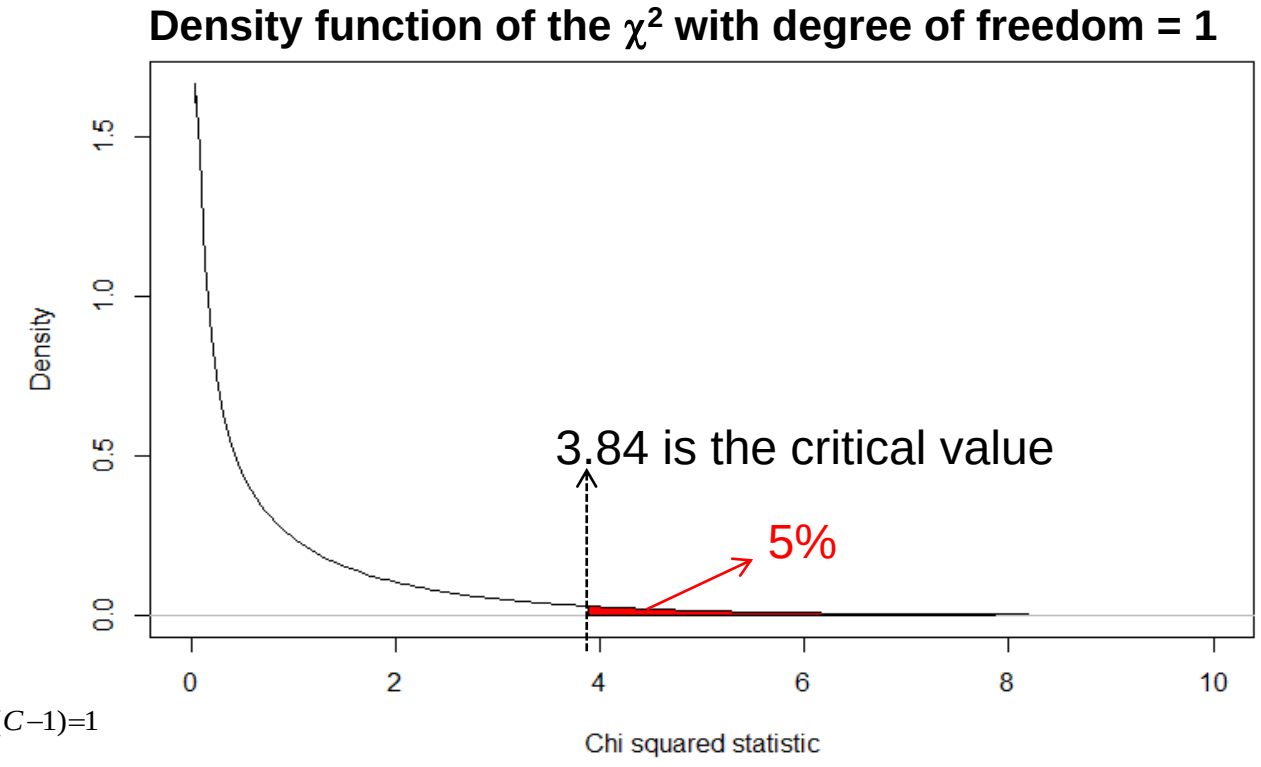
```
 Pearson's Chi-squared test with Yates' continuity correction

data:  my.tab
X-squared = 29.0595, df = 1, p-value = 7.019e-08
```

The χ^2 test for a 2x2 cross table

	diseased D	not diseased $\bar{D}$	
exposed E	$n_{11} = a$	$n_{12} = \mathrm{b}$	$n_E = n_{1\bullet}$
not exposed $\bar{E}$	$n_{21} = c$	$n_{22} = d$	$n_{\bar{E}} = n_{2\bullet}$
	$n_D = n_{\bullet 1}$	$n_{\bar{D}} = n_{\bullet 2}$	$\boldsymbol{n}$

Independence of row and column conditions can be tested with χ^2 test if the cell counts are sufficiently large:



$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \underset{\text{under } H_0}{\overset{a}{\sim}} \chi^2_{df=(R-1)\cdot(C-1)=1}$$

The approximation of the chi-square test is only valid, of number in all cells are >5.

Tests for row-column independence in cross tables
(for un-matched data)

- **Chi-Square Test [R×C tables]**

This test requires large sample sizes to be accurate.
An often quoted rule of thumb regarding sample size is that none of the expected cell values can be less than five.

$$\chi^2_P = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- **Yates' Continuity Corrected Chi-Square Test [2×2 tables]**

This test is similar to chi-square test above, but is adjusted for the continuity of the chi-square distribution.

$$\chi^2_Y = \sum_i \sum_j \frac{\left(max(0, |O_{ij} - E_{ij}| - 0.5)\right)^2}{E_{ij}}$$

- **Likelihood Ratio Test [R×C tables]**

Under independence the likelihood ratio statistic follows an asymptotic chi-square distribution.

$$\chi^2_{LR} = 2 \sum_i \sum_j O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

- **Fisher's Exact Test [2×2: hypergeometric, RxC: permutation distribution]**

Fisher's exact test is often used when sample sizes are small, but it is appropriate for all sample sizes.

Maternal drinking example: Chi-square test on reduced table

Maternal drinking and congenital malformations

Malformation	Alcohol consumption (average no. of drinks/day) 0	< 1	1–2	3–5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

Source: Graubard and Korn (1987).

Malformation	Drinks per day: < 1	≥ 1
absent	31530	951
present	86	7

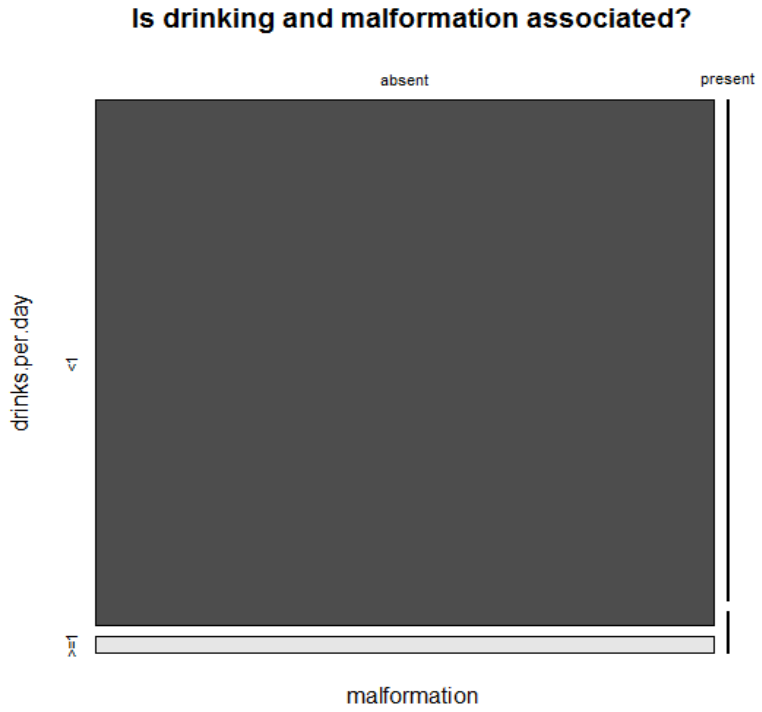
Since some cells have small counts, we can not apply the χ^2 test to the full table.

```
> my.tab <- as.table(rbind(c(17066+14464, 788+126+37),
+                          c(48+38, 5+1+1)))
> (Xsq <- chisq.test(my.tab))

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

data:  my.tab
X-squared = 5.3546, df = 1, p-value = 0.02067

Warning message:
In chisq.test(my.tab) : Chi-Quadrat-Approximation kann inkorrekt sein
```



Maternal drinking example: Fisher's Exact Test

Malformation	Drinks per day: < 1	≥ 1
absent	31530	951
present	86	7

Fisher's exact test for testing the null of independence of rows and columns in a contingency table with fixed marginals.

Do we meet this assumption??

```
> my.tab <- as.table(rbind(c(17066+14464, 788+126+37),
+                          c(48+38, 5+1+1)))
> fisher.test(my.tab)

 Fisher's Exact Test for Count Data

data:  my.tab
p-value = 0.01989
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.050591 5.823499
sample estimates:
odds ratio
  2.698369
```

Beside the p-value we get for a 2x2 table an estimate for OR (OR_D=OR_E) and with the CI a range of plausible values for the association.

Does Fisher's exact test work without all marginals are fixed? Without proof: the answer is yes

	D	$\bar{D}$	
E	$n_{11} = a$	$n_{12} = \mathrm{b}$	$n_E = n_{1\bullet}$
$\bar{E}$	$n_{21} = c$	$n_{22} = d$	$n_{\bar{E}} = n_{2\bullet}$
	$n_D = n_{\bullet 1}$	$n_{\bar{D}} = n_{\bullet 2}$	n

$$H_0: \quad p = P(E \mid D) = P(E \mid \bar{D})$$

We can use Fisher exact test without requiring that the marginals are fixed (without proof).

The Fisher exact test still leads a correct p-value of the 1-sided test and we restrict to 2x2 tables.

However, there are tests based on the Binomial distribution with larger power (D'Agostino et. al, 1988)

for derivations see e.g. "Opinionated lessons in statistics #33" by Prof. Bill Press:
http://granite.ices.utexas.edu/coursewiki/index.php/Segment_33._Contingency_Table_Protocols_and_Exact_Fisher_Test