Faculty of Health Sciences

Day 5: binary responses and 2×2 tables

Basic Statistics for health researchers

Alessandra Meddis

Section of Biostatistics, University of Copenhagen

November 8, 2023

Outline/Intended Learning Outcomes (ILOs)

Preliminaries

ILO: calculate 95% CIs for population proportions
ILO: distinguish between exact and approximate (asymptotic) 95% CIs

Group comparison

ILO: to define a suitable association measure and compute its 95% CI
ILO: to (correctly) use the χ^2 test and Fisher's test

Sample size and power calculation

ILO: to identify why and how to make power and sample size calculations
ILO: to analyse their strengths and limitations

Confounding

ILO: to exemplify confounding and its potential to be misleading
ILO: to name two commonly used remedies

Cohort vs case-control study

ILO: to differentiate the cohort and case-control designs
ILO: to restate which association measure(s) can be used for each design

Screening: jargon

ILO: to recognize some jargon

Paired binary data (if time allows)

ILO: to exemplify paired binary data
ILO: to calculate appropriate 95%-CI and p-values

Binary outcome

$$Y = \begin{cases} 1 & \text{event / positive / disease} \\ 0 & \text{no event / negative / non-disease} \end{cases}$$

Binary outcome

$$Y = \begin{cases} 1 & \text{event / positive / disease} \\ 0 & \text{no event / negative / non-disease} \end{cases}$$

Parameters

- Prevalence: proportion of the population with event at fixed time point.
  How many have the disease right now?
- Risk: probability that event occurs in given time period:
  How likely will a subject acquire the disease within 1-year?

Statistical inference

Estimating risks and prevalence

$$\widehat{p} = \text{Relative frequency} = \frac{\text{Number of events}}{\text{Number of subjects}} = \frac{x}{n}$$

Confidence limits: normal approximation (“large” n[1])

$$\left[\widehat{p} - 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}; \widehat{p} + 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

Confidence limits: “exact” (any n)

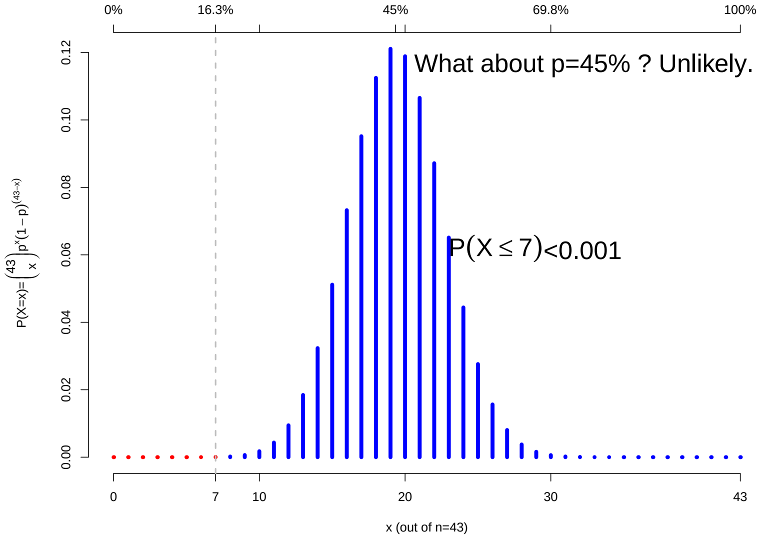
```
binom.test(x,n)
```

[1] rule of thumb: when both $x \geq 5$ and $n - x \geq 5$.

Exact confidence intervals (computation/intuition)

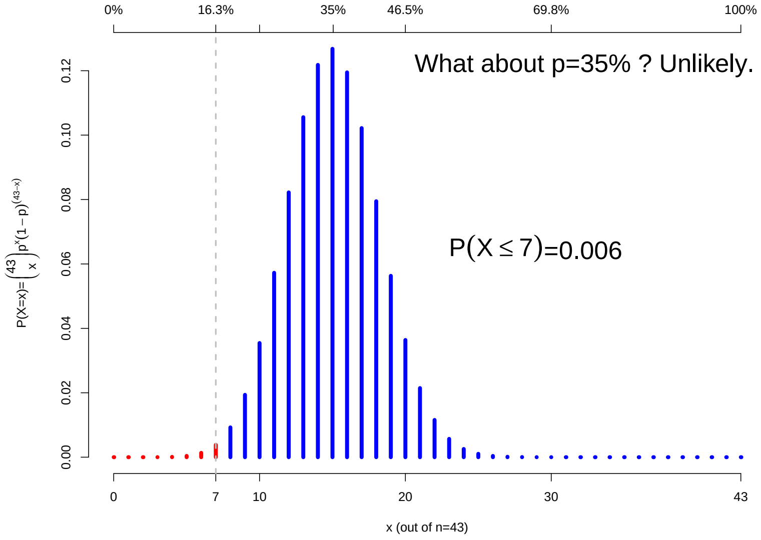
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [6.8; 30.7]$.

Exact confidence intervals (computation/intuition)



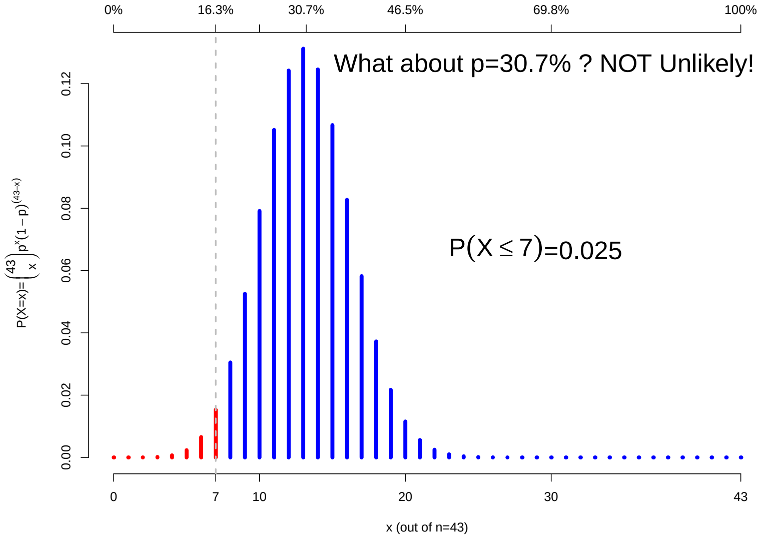
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [6.8; 30.7]$.

Exact confidence intervals (computation/intuition)



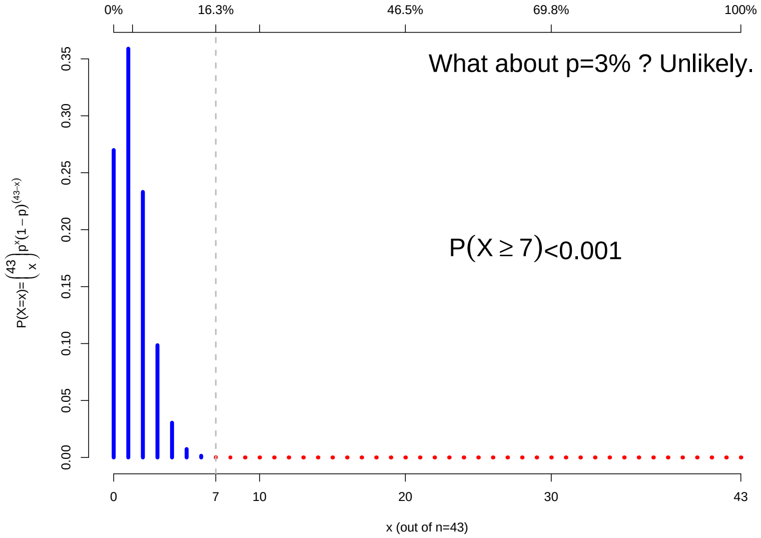
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [6.8; 30.7]$.

Exact confidence intervals (computation/intuition)



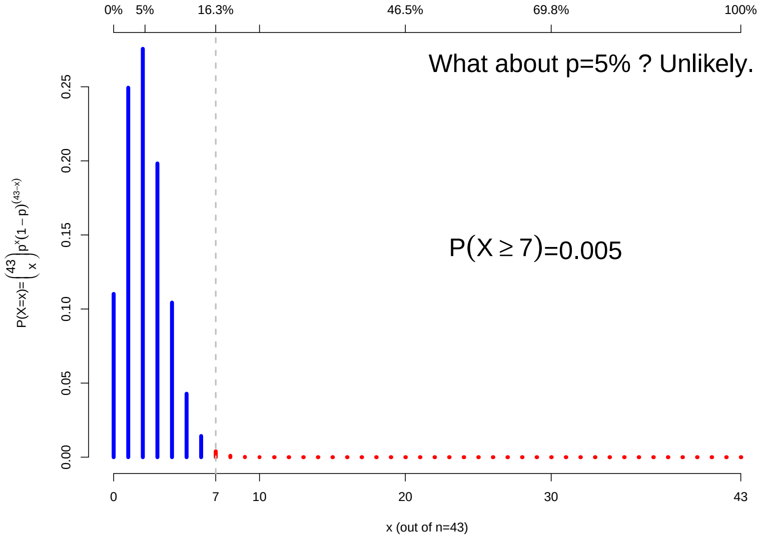
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [6.8; 30.7]$.

Exact confidence intervals (computation/intuition)



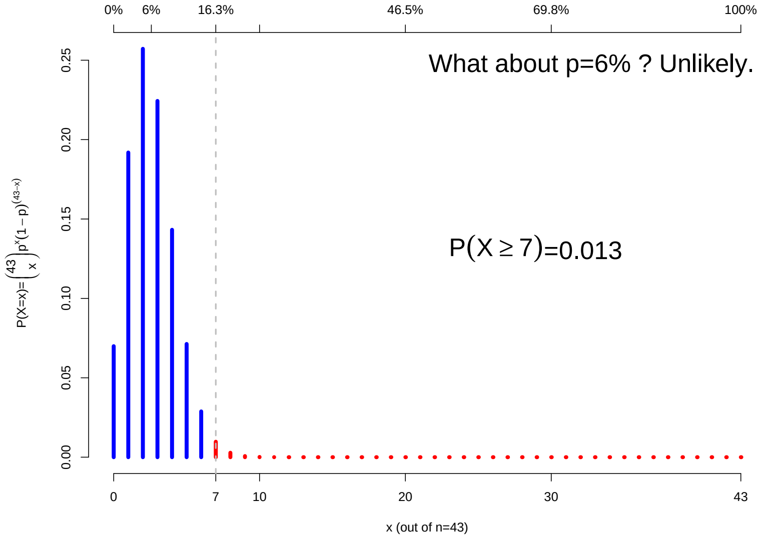
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [6.8; 30.7]$.

Exact confidence intervals (computation/intuition)



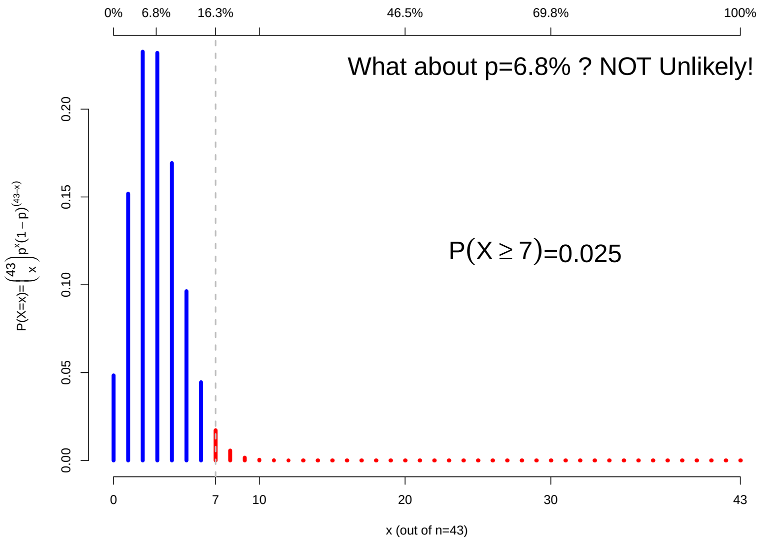
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [6.8; 30.7]$.

Exact confidence intervals (computation/intuition)



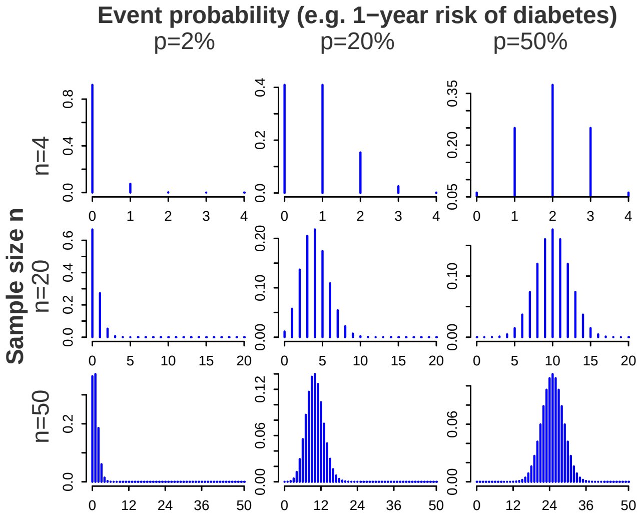
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [6.8; 30.7]$.

Exact confidence intervals (computation/intuition)



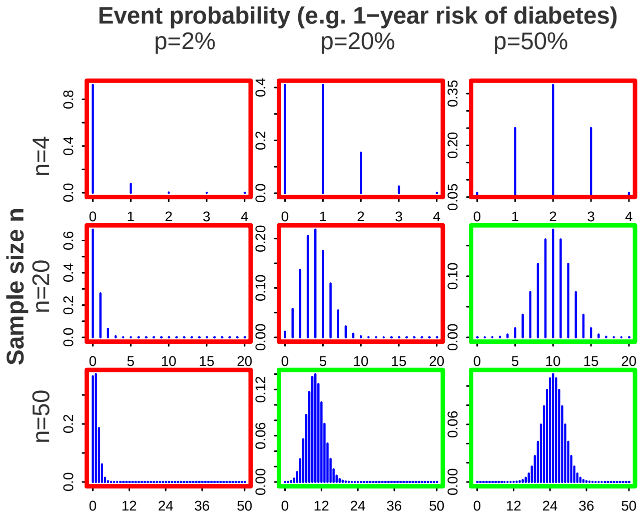
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [6.8; 30.7]$.

Normal approximation



- Binomial distribution: $P(X = x) = \binom{N}{x}p^x(1-p)^{N-x}$
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [5.2; 27.3]$.

Normal approximation



- "good"approximation if $np \geq 5$ and $n(1-p) \geq 5$ (green boxes).
- $x = 7$ and $n = 43$ leads to $\hat{p} = 16.3\%$ and 95% CI$= [5.2; 27.3]$.

Outline/Intended Learning Outcomes (ILOs)

Preliminaries

ILO: calculate 95% CIs for population proportions
ILO: distinguish between exact and approximate (asymptotic) 95% CIs

Group comparison

ILO: to define a suitable association measure and compute its 95% CI
ILO: to (correctly) use the χ^2 test and Fisher's test

Sample size and power calculation

ILO: to identify why and how to make power and sample size calculations
ILO: to analyse their strengths and limitations

Confounding

ILO: to exemplify confounding and its potential to be misleading
ILO: to name two commonly used remedies

Cohort vs case-control study

ILO: to differentiate the cohort and case-control designs
ILO: to restate which association measure(s) can be used for each design

Screening: jargon

ILO: to recognize some jargon

Paired binary data (if time allows)

ILO: to exemplify paired binary data
ILO: to calculate appropriate 95%-CI and p-values

Case: clinical trial on Dalteparin [3]

Data: $n = 85$ diabetic patients with peripheral arterial occlusive disease and chronic foot ulcers, randmomized (double-blind) to:

- Placebo $(n = 42)$
- Dalteparin $(n = 43)$

Outcome:

Category [2]	Label
intact skin	healed
decreased ulcer area $\geq 50\%$	improved
increased ulcer area $\geq 50\%$	impaired
decreased or increased ulcer area $< 50\%$	unchanged
amputation above/below ankle	amputation

Research question: Does Dalteparin improve the outcome, when injected once daily until ulcer healing or for a maximum of 6 months?

[2] mutually exclusive.

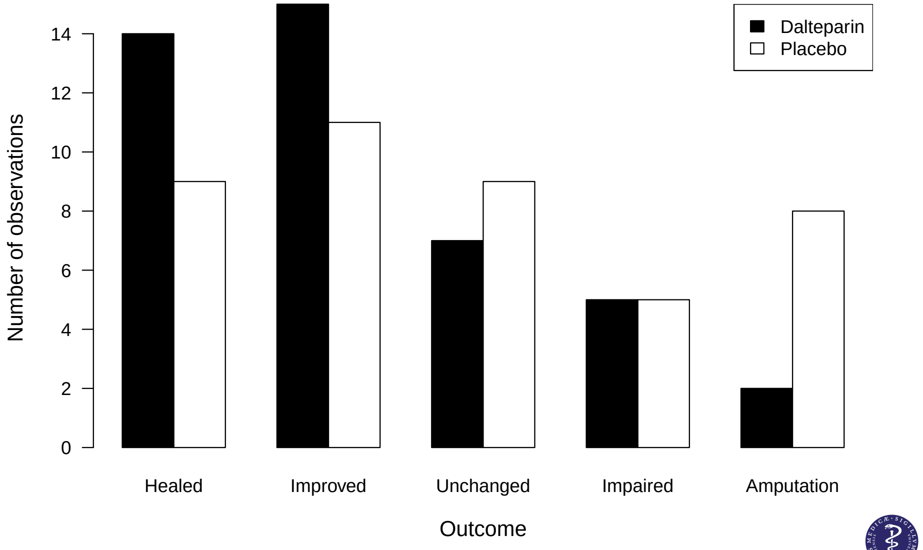
[3] Kalani et al. *Diabetes Care* **26**: 2575-2580, 2003

Frequency table

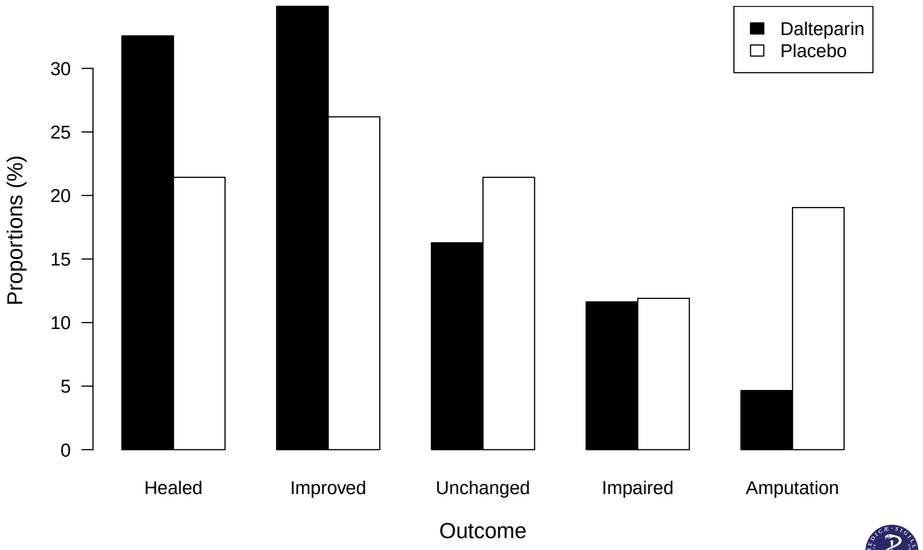
	Dalteparin	Placebo
Healed	14 (33%)	9 (21%)
Improved	15 (35%)	11 (26%)
Unchanged	7 (16%)	9 (21%)
Impaired	5 (12%)	5 (12%)
Amputation	2 (5%)	8 (19%)
total (100%)	43	42

- Summarizes the outcome data.
- Prepare/Format data for analyzes.

Barplot (frequencies)



Barplot (proportions[4])



[4] often better when sample sizes are not equal in both groups.

Here we pool the outcome categories as follows

Category	Dichotomized outcome
intact skin	better
ulcer area decreased $\geq 50\%$	
decreased or increased ulcer area $< 50\%$	worse
increased ulcer area $\geq 50\%$	
amputation above/below ankle	

Important: this dichotomization should be prespecified (i.e. decision made before seeing the data). [5]

[5] For an illustration of why prespecification matters, see e.g. Austin & Goldwasser. "Pisces did not have increased heart failure: data-driven comparisons of binary proportions between levels of a categorical variable can result in incorrect statistical significance levels." Journal of clinical epidemiology 61.3 (2008): 295-300.

Group comparison

Placebo group

$$\text{Risk of worse outcome} = \frac{22}{42} = \widehat{p}_1$$

Dalteparin group

$$\text{Risk of worse outcome} = \frac{14}{43} = \widehat{p}_2$$

[6] whenever possible, we prefer using risk ratios or risk differences to odds ratios. They are often better understood and easier to communicate!

Group comparison

Placebo group

$$\text{Risk of worse outcome} = \frac{22}{42} = \widehat{p}_1$$

Dalteparin group

$$\text{Risk of worse outcome} = \frac{14}{43} = \widehat{p}_2$$

Association measures[6]

Relative risk: $\frac{\widehat{p}_1}{\widehat{p}_2}$ Odds ratio: $\frac{\frac{\widehat{p}_1}{1-\widehat{p}_1}}{\frac{\widehat{p}_2}{1-\widehat{p}_2}}$ Risk difference: $\widehat{p}_1 - \widehat{p}_2$

[6] whenever possible, we prefer using risk ratios or risk differences to odds ratios. They are often better understood and easier to communicate!

2x2 contingency table

		Response		
		yes	no	total
Exposure	yes	a	b	a+b
	no	c	d	c+d
	total	a+c	b+d	N

Risk estimates

$$\widehat{p}_1 = \frac{a}{a+b} \qquad \widehat{p}_2 = \frac{c}{c+d}$$

Relative risk

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)}$$

		Response		
		yes	no	total
Exposure	yes	a	b	a+b
	no	c	d	c+d
	total	a+c	b+d	N

Standard error of $\log(\widehat{RR})$ and confidence interval of RR [7]

$$\widehat{\sigma} = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

$$CI_{95\%} = \left[\widehat{RR} \cdot \exp(-1.96\,\widehat{\sigma}) \;;\; \widehat{RR} \cdot \exp(1.96\,\widehat{\sigma})\right]$$

[7] This method is "good enough" with "large enough" sample sizes.

Relative risk: placebo versus dalteparin

$$\widehat{RR} = \frac{22/42}{14/43} = 1.609$$

		Outcome		
		worse	better	total
Treatment	placebo	22	20	42
	dalteparin	14	29	43
	total	36	49	85

Standard error of $\log(\widehat{RR})$ and confidence interval

$$\hat{\sigma} = \sqrt{\frac{1}{22} - \frac{1}{42} + \frac{1}{14} - \frac{1}{43}} = 0.264$$

$$CI_{95\%} = [0.959; 2.7]$$ (does include 1)

Risk difference

$$\widehat{\Delta} = \frac{a}{a+b} - \frac{c}{c+d}$$

		Response		
		yes	no	total
Exposure	yes	a	b	a+b
	no	c	d	c+d
	total	a+c	b+d	N

Standard error of $\widehat{\Delta}$ and confidence interval [8]

$$\widehat{\sigma} = \sqrt{ab/(a+b)^3 + cd/(c+d)^3}$$

$$CI_{95\%} = \left[\widehat{\Delta} - 1.96\,\widehat{\sigma} \;;\; \widehat{\Delta} - 1.96\,\widehat{\sigma}\right]$$

[8] This method is "good enough" with "large enough" sample sizes.

Risk difference: placebo versus dalteparin

$$\widehat{\Delta} = \frac{22}{42} - \frac{14}{43} = 0.198$$

		Outcome		
		worse	better	total
Treatment	placebo	22	20	42
	dalteparin	14	29	43
	total	36	49	85

Standard error of $\widehat{\Delta}$ and confidence interval

$$\widehat{\sigma} = \sqrt{22 \cdot 20/42^3 + 14 \cdot 29/43^3} = 0.105$$

$$CI_{95\%} = [-0.008 \;;\; 0.404]$$ (does include 0)

Odds Ratio (OR)

Concept **needed** for

- case-control studies
- logistic regression

Odds: are defined as "risk of event divided by risk of no event"

$$\boxed{\text{odds} = p/(1-p)}\ ,$$

and the risk can be computed back from the odds, $p = \text{odds}/(1+\text{odds})$. Odds are difficult to interpret, but if risks are small, then risks $\approx$ odds.

The Odds ratio (OR) is defined as the ratio of the odds,

$$\boxed{OR = \frac{\text{odds}_1}{\text{odds}_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}}\ .$$

OR are difficult to interpret, but from the equation...

$$RR = \frac{OR}{\left\{1-p_2\right\} + p_2 OR},$$

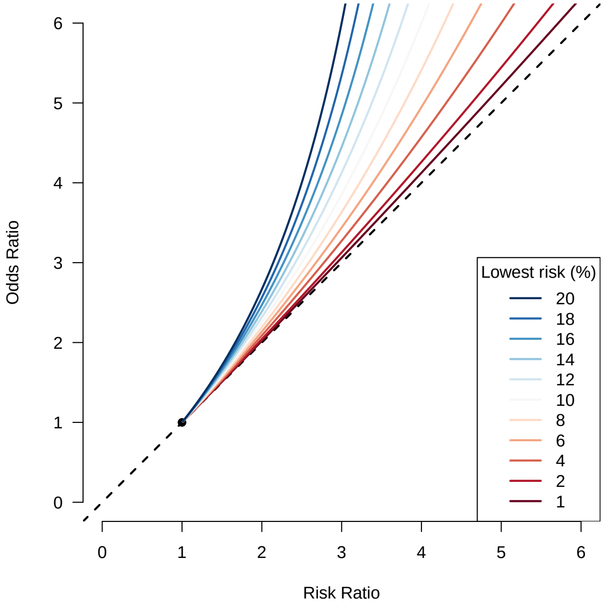
...we can first conclude:

- $OR > 1 \Leftrightarrow RR > 1$
- $OR = 1 \Leftrightarrow RR = 1$
- $OR < 1 \Leftrightarrow RR < 1$

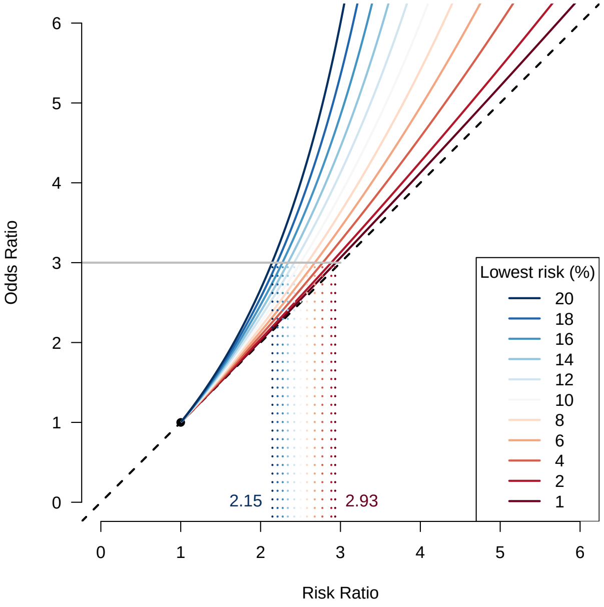
...and further conclude that

- the OR is sufficient to deduce whether a risk increases or decreases.
- if p_2 is small (e.g. rare disease), then $OR \approx RR$.

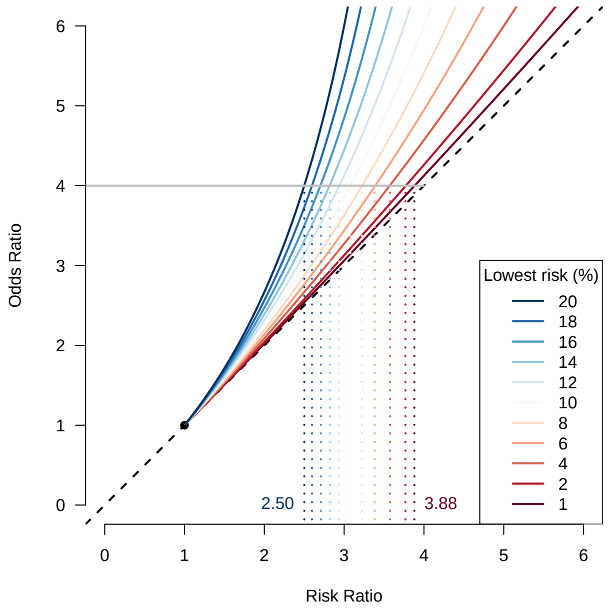
When is $OR \approx RR$?



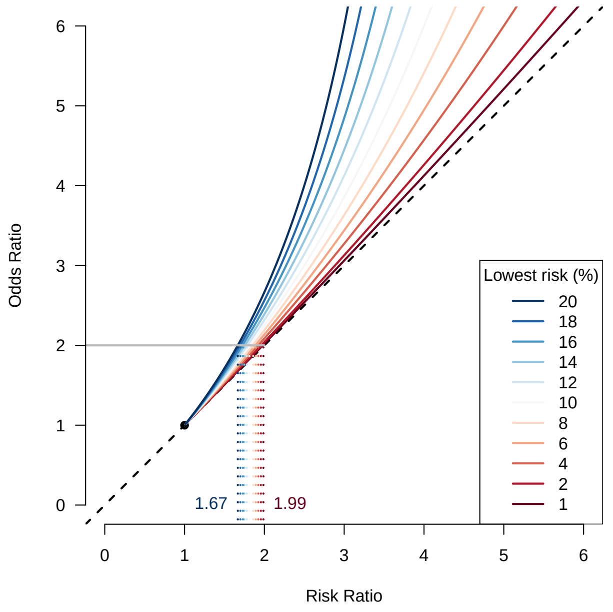
When is $OR \approx RR$?



When is $OR \approx RR$?



When is $OR \approx RR$?



Odds ratio

$$\widehat{OR} = \frac{\frac{a/(a+b)}{b/(a+b)}}{\frac{c/(c+d)}{d/(c+d)}} = \frac{a \cdot d}{b \cdot c}$$

Exposure / Response

		yes	no	total
Exposure	yes	a	b	a+b
	no	c	d	c+d
	total	a+c	b+d	N

Standard error of $\log(\widehat{OR})$ and confidence interval[9]

$$\widehat{\sigma} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$CI_{95\%} = \left[\widehat{OR} \cdot \exp(-1.96\,\widehat{\sigma}); \widehat{OR} \cdot \exp(1.96\,\widehat{\sigma})\right]$$

[9] This method is “good enough” with “large enough” sample sizes.

Odds ratio: placebo versus dalteparin

$$\widehat{OR} = \frac{22 \cdot 29}{14 \cdot 20} = 2.279$$

Treatment / Outcome

		worse	better	total
Treatment	placebo	22	20	42
	dalteparin	14	29	43
	total	36	49	85

Standard error of $\log(\widehat{OR})$ and confidence interval

$$\widehat{\sigma} = \sqrt{\frac{1}{22} + \frac{1}{20} + \frac{1}{14} + \frac{1}{29}} = 0.449$$

$CI_{95\%} = [0.946; 5.491]$ (does include 1)

Reporting results

The relative risk (of worsening) of group 1 (Dalteparin) versus group 2 (Placebo) is estimated as

$$RR = \frac{14/43}{22/42} = 0.622$$

Equivalent statements:

- *The risk in group 1 is reduced by a factor 0.622 compared to group 2.*
- *The risk in group 1 is* **37.8% lower** *than in group 2.*[10]
- *The risk in group 2 is 1.609 times higher than in group 1.*[11]
- *The risk in group 2 is* **60.9% higher** *than in group 1.*

[10] because 1-0.622=0.378

[11] because 1/0.622=1.609

Testing independence in a randomized clinical trial

Null hypothesis H_0: the treatment has no effect.

$$\begin{aligned}
&\text{Prob(worse given dalteparin)} = \text{Prob(worse given placebo)} \\
\Leftrightarrow \quad & p_1 - p_2 = 0 \qquad \text{(Difference =0)} \\
\Leftrightarrow \quad & \frac{p_1}{p_2} = 1 \qquad \text{(Relative risk =1)} \\
\Leftrightarrow \quad & \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = 1 \qquad \text{(Odds ratio =1)}
\end{aligned}$$

Popular tests of independence between the treatment group and the outcome groups:

- χ^2 test (normal approximation)[12]
- Fisher's exact test: recommended as the default choice! [13]

[12] This method is "good enough" with "large enough" sample sizes.

[13] Recommended because: Why approximate when you can get the exact?

The χ^2 test statistic

$$\chi^2 = \sum \frac{(\text{observed counts} - \text{expected counts})^2}{\text{expected counts}}$$

Observed counts

		Response		
		yes	no	total
Exposure	yes	a	b	a+b
	no	c	d	c+d
	total	a+c	b+d	N

Expected counts

		Response		
		yes	no	total
Exposure	yes	(a+b)(a+c)/N	(a+b)(b+d)/N	a+b
	no	(c+d)(a+c)/N	(c+d)(b+d)/N	c+d
	total	a+c	b+d	N

- under the null hypothesis the groups are identical, hence data can be merged into a single group
- in a population of size n, for a given risk of event p, we expect to see (on average) np events in this population

The expected counts are calculated under the null hypothesis.

Rule of thumb: a valid analysis requires that all expected counts are ≥ 5.

Test results

Null hypothesis:
dalteparin treatment has no effect for chronic foot ulcers.

Test	p-value
Fisher's exact test	0.0808
Pearson's χ^2 test	0.0644
Pearson's χ^2 test with Yates' continuity correction[14]	0.1032

R code:

```
tab <- rbind(c(22,20),c(14,29))
fisher.test(tab)                 # always works (default choice!)
chisq.test(tab,correct=FALSE)    # fine with large samples
chisq.test(tab,correct=TRUE)     # no longer useful
```

[14] Expected to be more precise than the usual Pearson's χ^2 test when the sample size is very small. **NOT RECOMMENDED**, with small sample sizes, use Fisher's test instead.

A note of caution

Because the (simple) formulas for the 95% CI (of the previous slides) are based on large sample size approximations, they are not necessarily consistent with the result of the Fisher's exact test, especially with "very small" sample sizes.

Example:

	event	no event
exposed	5	12
non-exposed	8	3

- $\widehat{p}_1 = 8/11 = 0.73, \quad \widehat{p}_2 = 5/17 = 0.29.$
- $\widehat{\Delta}$= 0.43 (0.09 ; 0.77)
- $\widehat{RR}$= 2.47 (1.09 ; 5.62)
- $\widehat{OR}$= 6.40 (1.18 ; 34.61)
- p-values from Fisher's exact test and Pearson's χ^2 (with and without Yates correction) are 0.051, 0.063 and 0.025, respectively.

Here the confidence intervals show a significant result, but not Fisher's test.

Advanced methods and software[15] are available to avoid running into this kind of inconsistency between hypothesis test and confidence intervals.

Fortunately, it is rare that we run into this problem....
and even rarer that it matters for the interpretation.

[15] see R package exact2x2 and references in the help documentation.

Larger contigency tables (1/2)

If the table is not 2x2 but, e.g., 3x4 or 2x4, the χ^2 test and Fisher's exact test are testing an "ANOVA-like" null hypothesis similarly to what the F-test does to compare several means.

First example:

	underweight	normal	overweight	obese
no SCD	9	51	20	8
SCD	23	61	3	1

R code:

```
fisher.test(table(d$SCD,d$BMIgroup))
```

returns a p-value <0.001, for the null hypothesis

H$_0$: "the prevalence of SCD is the same in all groups of BMI"

that is, "no association between BMI group and SCD".

Larger contigency tables (2/2)

Second example:

	underweight	normal	overweight	obese
age=$[16, 25)$	14	45	1	1
$[25, 30)$	3	25	3	1
$[30, 67]$	15	42	19	7

R code:

```
fisher.test(table(d$ageGroup,d$BMIgroup))
```

returns p-value=0.004, for the null hypothesis

H_0: "the prevalence of each BMI group is the same in all groups of age "

that is, "no association between BMI group and age".

Outline/Intended Learning Outcomes (ILOs)

Preliminaries

ILO: calculate 95% CIs for population proportions

ILO: distinguish between exact and approximate (asymptotic) 95% CIs

Group comparison

ILO: to define a suitable association measure and compute its 95% CI

ILO: to (correctly) use the χ^2 test and Fisher's test

Sample size and power calculation

ILO: to identify why and how to make power and sample size calculations

ILO: to analyse their strengths and limitations

Confounding

ILO: to exemplify confounding and its potential to be misleading

ILO: to name two commonly used remedies

Cohort vs case-control study

ILO: to differentiate the cohort and case-control designs

ILO: to restate which association measure(s) can be used for each design

Screening: jargon

ILO: to recognize some jargon

Paired binary data (if time allows)

ILO: to exemplify paired binary data

ILO: to calculate appropriate 95%-CI and p-values

Textbook formula ("large n" approximation)

Sample size and power calculation is mostly useful for designing clinical trials. However, this could be a useful tool in observational studies to understand what is possible to achieve with the available data.

When calculating the sample size we need to specify:

- expected p_1, p_2
- the desired power $(1-\beta)$ and Type I error (α)

$$n = \frac{\left\{ z_{\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_\beta\sqrt{p_1(1-p_1)+p_2(1-p_2)} \right\}^2}{(p_1-p_2)^2}$$

- z_γ is the γ-quantile of a standard normal distribution [16]
- $\bar{p} = (p_1+p_2)/2$.
- n: number of observations in **each** group.

Reverse the formula to compute:

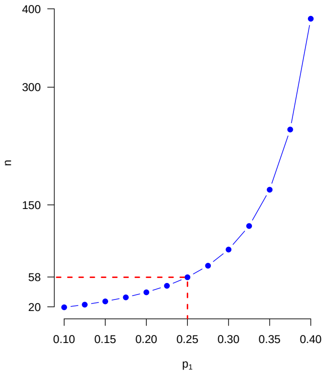
- Power for a given sample size: for expected values of p_1 and p_2 and desired n and α.
- Least detectable difference (or ratio): $\delta = p_1 - p_2$ (or $r = p_1/p_2$) for given n, expected p_1, desired α and minimal power $(1-\beta)$.

[16] $z_{\alpha/2} = -1.96$ for $\alpha = 5\%$ and $z_\beta = 0.84$ is 1.28 for $1-\beta = 80\%$

Sample size calculation

Standard software can be used, e.g. R:

```
power.prop.test(p1 = 0.25, p2 = 0.5, power=0.8)

     Two-sample comparison of proportions power calculation

              n = 57.67344
             p1 = 0.25
             p2 = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

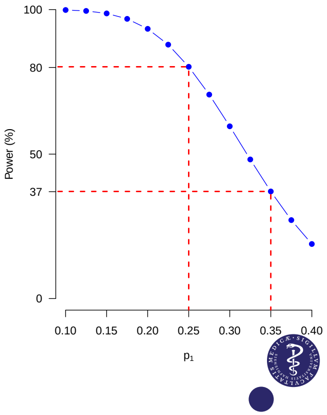


- $n = 58$ subjects needed in **each** group (i.e. 116 in total) to detect significant risk difference with a power of 80%, if the risks in the two groups are 25% and 50%.

Power calculation

Example: an initial calculation suggests $n = 58$ subjects per group (i.e. 116 in total), for detecting a difference of 25% survival between the two groups, assuming 50% survival in the placebo group (with 80% power). But what does the power become if we were too optimistic with the expected treatment effect? E.g. what if the difference in survival probability is only 15%?

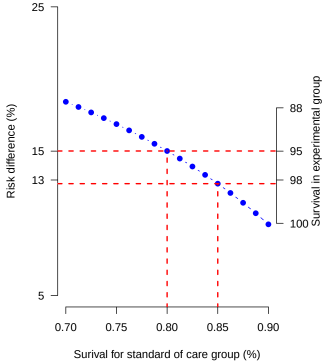
```
power.prop.test(n=58, p1 = 0.35, p2 = 0.5)

     Two-sample comparison of proportions power calculation

              n = 58
             p1 = 0.35
             p2 = 0.5
      sig.level = 0.05
          power = 0.3707966
    alternative = two.sided

NOTE: n is number in *each* group
```



Least detectable difference

Example: My grant can finance a total sample size of $n = 150$ (i.e. 75 per group). What is the smallest survival difference that I can hope to show with a decent power (e.g. 80%), if I expect 80% survival in the "standard of care" (i.e. control) group? And if I expect 85% in the "standard of care" group?

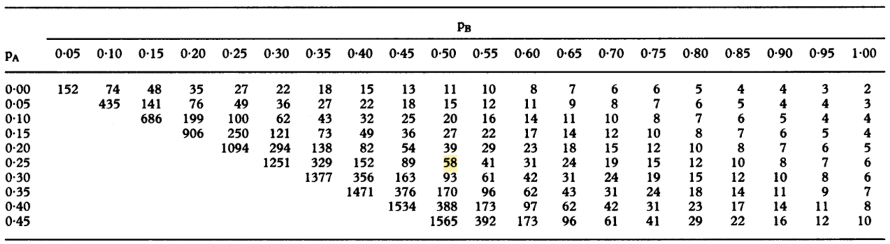
```
power.prop.test(n=75, p1 = 0.8, power=0.8)

     Two-sample comparison of proportions power calculation

              n = 75
             p1 = 0.8
             p2 = 0.950095
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```



Note: you need to supply a value for p1, not p2, otherwise the software is looking for a lower risk and it returns 0.72.

Digression: Tables also exist (for sample size calculation)[17]

TABLE II—*Sample sizes to detect a difference in two proportions, p_A and p_B, at a 5% significance level with 80% power*



	p_B																			
p_A	0·05	0·10	0·15	0·20	0·25	0·30	0·35	0·40	0·45	0·50	0·55	0·60	0·65	0·70	0·75	0·80	0·85	0·90	0·95	1·00
0·00	152	74	48	35	27	22	18	15	13	11	10	8	7	6	6	5	4	4	3	2
0·05		435	141	76	49	36	27	22	18	15	12	11	9	8	7	6	5	4	4	3
0·10			686	199	100	62	43	32	25	20	16	14	11	10	8	7	6	5	4	4
0·15				906	250	121	73	49	36	27	22	17	14	12	10	8	7	6	5	4
0·20					1094	294	138	82	54	39	29	23	18	15	12	10	8	7	6	5
0·25						1251	329	152	89	58	41	31	24	19	15	12	10	8	7	6
0·30							1377	356	163	93	61	42	31	24	19	15	12	10	8	6
0·35								1471	376	170	96	62	43	31	24	18	14	11	9	7
0·40									1534	388	173	97	62	42	31	23	17	14	11	8
0·45										1565	392	173	96	61	41	29	22	16	12	10

- Here again we can see again $n = 58$ (as in a previous slide).

[17] **Source:** Campbell, Julious & Altman (1995). Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ, 311(7013), 1145-1148.

Outline/Intended Learning Outcomes (ILOs)

Preliminaries
- ILO: calculate 95% CIs for population proportions
- ILO: distinguish between exact and approximate (asymptotic) 95% CIs

Group comparison
- ILO: to define a suitable association measure and compute its 95% CI
- ILO: to (correctly) use the χ^2 test and Fisher's test

Sample size and power calculation
- ILO: to identify why and how to make power and sample size calculations
- ILO: to analyse their strengths and limitations

Confounding
- ILO: to exemplify confounding and its potential to be misleading
- ILO: to name two commonly used remedies

Cohort vs case-control study
- ILO: to differentiate the cohort and case-control designs
- ILO: to restate which association measure(s) can be used for each design

Screening: jargon
- ILO: to recognize some jargon

Paired binary data (if time allows)
- ILO: to exemplify paired binary data
- ILO: to calculate appropriate 95%-CI and p-values

Confounding

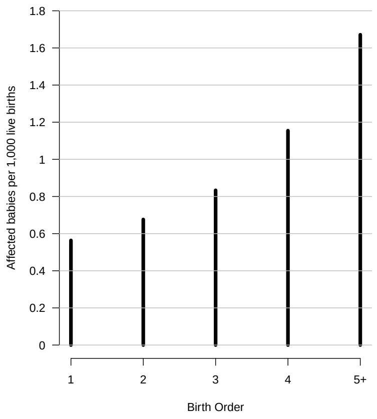
"A simple definition of confounding is the confusion of effects. This definition implies that the effect of the exposure is mixed with the effect of another variable, leading to a bias."[18]

Failing to take a confounding variable into account can lead to a **false conclusion** that the outcome are in a **causal relationship** with the predictor variable.

Confounding variables are typically encountered in observational studies, but not in "ideal" randomized experiments.

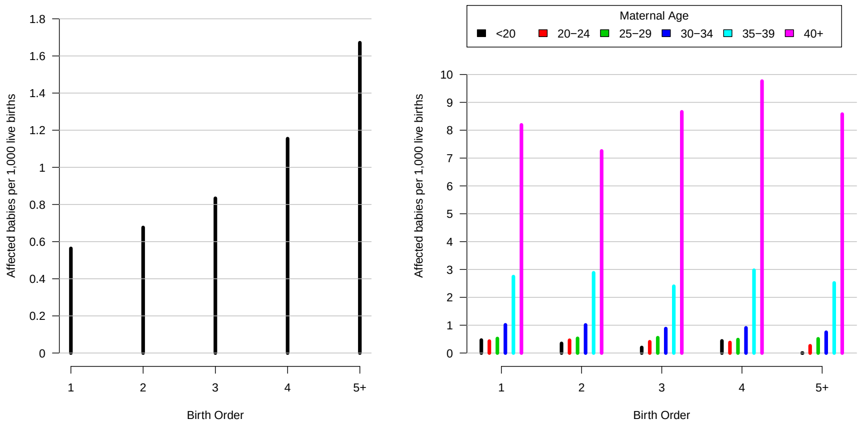
[18] Rothman (2012), Epidemiology: an introduction.

Confounding example (birth order and risk of Down syndrome [19])



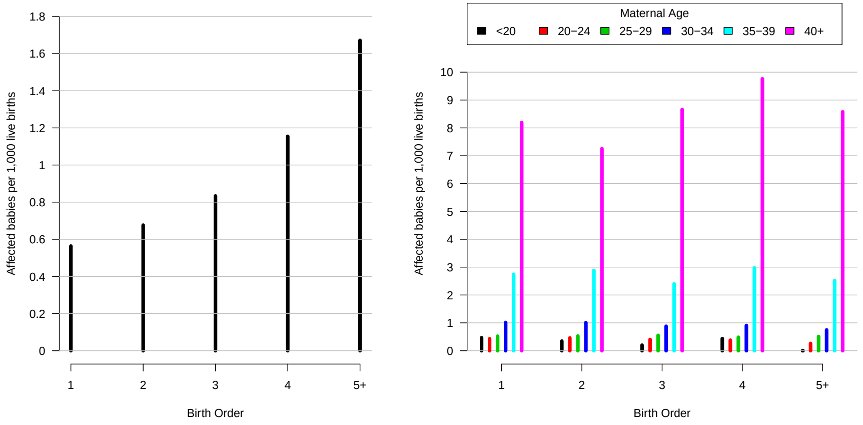
[19] Stark and Mantel (1966), J. Natl. Cancer Inst. 37(5) 687–698.

Confounding example (birth order and risk of Down syndrome [19])

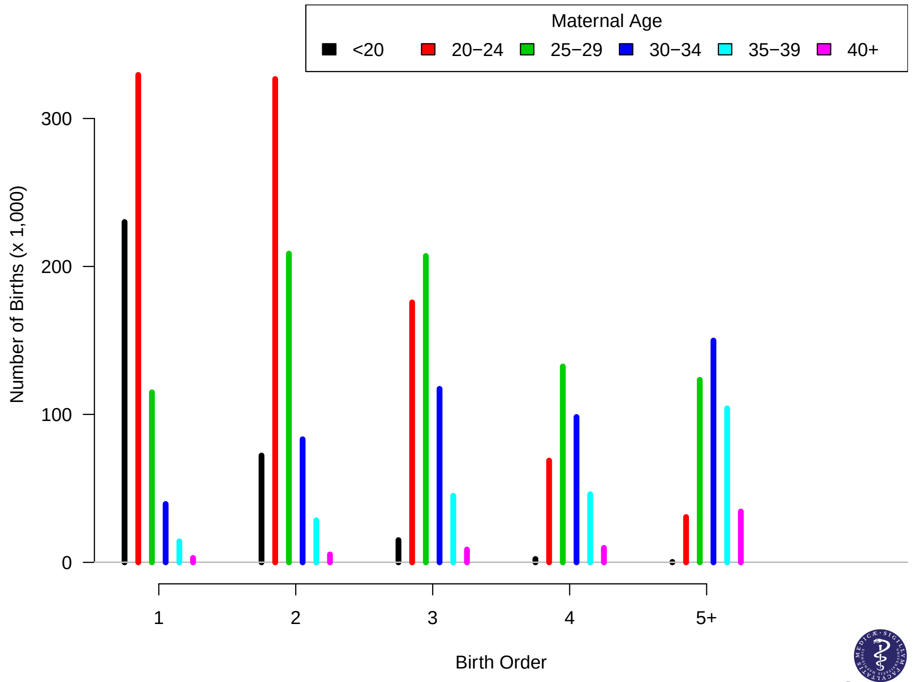


[19] Stark and Mantel (1966), J. Natl. Cancer Inst. 37(5) 687–698.

Confounding example (birth order and risk of Down syndrome [19])



[19] Stark and Mantel (1966), J. Natl. Cancer Inst. 37(5) 687–698.



When can association mean causation? (1/2)

We usually say that (statistical) association does not imply causation.

In presence of confounding we might not be able to identify the true causal effect.

We need (among others) that the groups we are comparing are similar with respect to everything except the treatment under study (exchangeability assumption).

When we succeed to correctly control for confounding, conditional exchangeability holds and association can be interpreted as causation.

When can association mean causation? (2/2)

An example where association implies causation is “ideal” randomized experiments.
The randomization ensures that the two groups that we compare are similar with respect to everything except the intervention / treatment under study. Hence, if a difference in outcome is observed between the two groups, then we can be confident that this is the consequence of this unique difference in exposure / treatment.

In non-randomized (or non “ideally” randomized) experiments the two compared groups will usually differ with respect to more than one characteristic. This generates multiple plausible explanations for the observation of the difference in outcome – some causal and some non causal.

Adjusted analysis

Suppose that in addition to the outcome and the exposure group a categorical confounder variable (e.g. gender) is measured for each individual.

- Subgroup analysis

Analyze 2x2 contingency tables separately in each strata defined by the confounder variable.

- Logistic regression (see Lecture 6)

To compute a “weighted” average of the subgroup analyses, assuming that the exposure-outcome association is the same in all subgroups.[20].

[20] Applicable also with continuous confounders.

Outline/Intended Learning Outcomes (ILOs)

Preliminaries
- ILO: calculate 95% CIs for population proportions
- ILO: distinguish between exact and approximate (asymptotic) 95% CIs

Group comparison
- ILO: to define a suitable association measure and compute its 95% CI
- ILO: to (correctly) use the χ^2 test and Fisher's test

Sample size and power calculation
- ILO: to identify why and how to make power and sample size calculations
- ILO: to analyse their strengths and limitations

Confounding
- ILO: to exemplify confounding and its potential to be misleading
- ILO: to name two commonly used remedies

Cohort vs case-control study
- ILO: to differentiate the cohort and case-control designs
- ILO: to restate which association measure(s) can be used for each design

Screening: jargon
- ILO: to recognize some jargon

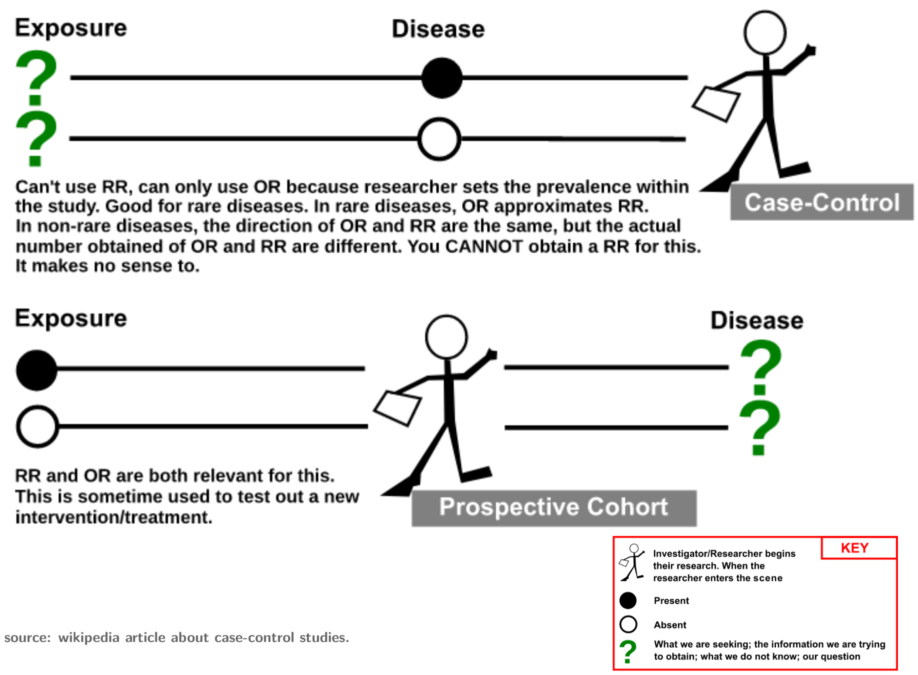
Paired binary data (if time allows)
- ILO: to exemplify paired binary data
- ILO: to calculate appropriate 95%-CI and p-values

Observational study design

In a prospective **cohort study**, an outcome or disease-free study population is first identified by an exposure (e.g., onset of diabetes) or other inclusion criteria and followed in time until the disease or outcome of interest occurs.

Case-control studies identify subjects by outcome status at the outset of the investigation. First, subjects with outcome are identified and classified as cases. For each case a given number of controls (e.g., 4) are selected. A candidate control is a subject without the outcome but from the same source population.

Observational Study Designs: Case Control vs Cohort



Can't use RR, can only use OR because researcher sets the prevalence within the study. Good for rare diseases. In rare diseases, OR approximates RR. In non-rare diseases, the direction of OR and RR are the same, but the actual number obtained of OR and RR are different. You CANNOT obtain a RR for this. It makes no sense to.

RR and OR are both relevant for this. This is sometime used to test out a new intervention/treatment.

Investigator/Researcher begins their research. When the researcher enters the scene

● Present

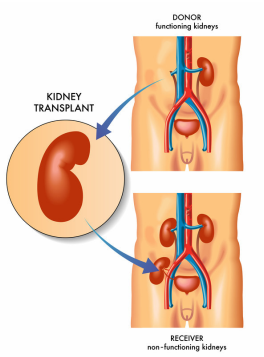
○ Absent

? What we are seeking; the information we are trying to obtain; what we do not know; our question

source: wikipedia article about case-control studies.

Cohort study: example from Egerup et al. (2020) [21]

Research question: How larger is the 1-year risk of infection (leading to an hospitalization) among newborns of kidney-transplanted women?



		Infection within first year of life		
		yes	no	total
Kidney-transplanted mother	yes	26	98	124
	no	133	1098	1231
	total	159	1196	1355

The estimated risk ratio is $\widehat{RR} = 1.94$ ($\text{CI}_{95\%} = [1.33; 2.83]$).

[21] Egerup et al. "Increased risk of neonatal complications and infections in children of kidney-transplanted women: A nationwide controlled cohort study." American Journal of Transplantation (2020).

Case-control study: example of Frachon et al. [22]

Research question: Is the use of benfluorex associated with unexplained mitral regurgitation?

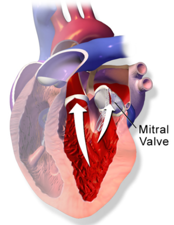
Mediator : dès 1998, l'Assu. mettait en garde sur les risq
Le Mediator serait responsable de 500 à 2000 décès en France
Chez Servier, on l'appelait le « Mediator »
Les patients ayant pris du Mediator doivent consulter
Mediator
Comment Servier a fait corriger le ra

- Case study described in the movie "150 Milligrams" (2016) (The original title in French is "La fille de Brest")
- France's biggest modern health scandal

[22] Frachon et al. "Benfluorex and unexplained valvular heart disease: a case-control study." PloS one 5.4 (2010).

Case-control study: example of Frachon et al.[23]

		“unexplained” mitral regurgitation		
		yes	no	total
Benfluorex use	yes	19	3	24
	no	8	51	59
	total	27	54	81



Mitral Valve Regurgitation

$\widehat{OR} = 40.4\ (CI_{95\%} : [9.7; 168])$

The number of controls (here 2 per case) is defined by the study design. Hence we cannot estimate risks as one minus the proportions of controls among exposed and non-exposed...

- The statistic $\widehat{RR}$ depends also on the ratio between controls and cases and should **not** be used for measuring association in case-control studies.
- The statistic $\widehat{OR}$ works.

[23] Frachon et al. "Benfluorex and unexplained valvular heart disease: a case-control study." PloS one 5.4 (2010).

Why does $\widehat{OR}$ work? (1/2)

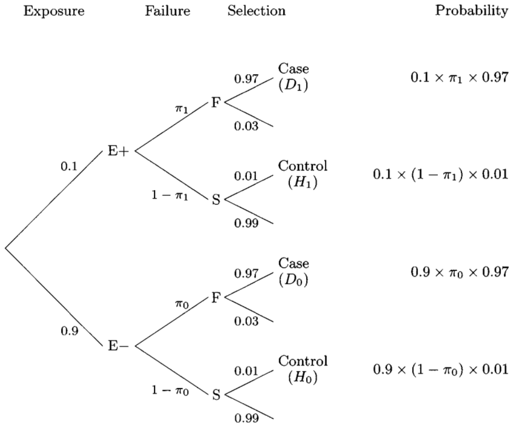


Fig. 16.1. The probability model in the study base.

$$OR = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

- 97% of the cases are included in the case-control study and 1% of the “non cases” are selected as controls; all included “blinded” from exposure (i.e. before looking for the information on the exposure).
- Connection to notations of previous slides $\pi_1 = p_1$ and $\pi_0 = p_2$.
- E=”exposure”, F=”Fail”, S=”Survive”, D=”Disease”, H=”Healthy”.
- source: “Statistical models in Epidemiology”, by Clayton and Hills, page 155.

Why does $\widehat{OR}$ work? (2/2)

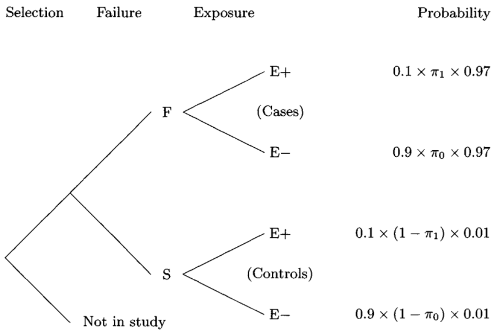

Selection
Failure
Exposure
Probability
F
(Cases)
E+
$0.1 \times \pi_1 \times 0.97$
E−
$0.9 \times \pi_0 \times 0.97$
S
(Controls)
E+
$0.1 \times (1-\pi_1) \times 0.01$
E−
$0.9 \times (1-\pi_0) \times 0.01$
Not in study

Fig. 16.2. The probability tree for the retrospective argument.

$$\widehat{OR} \approx \frac{\frac{0.1\times\pi_1\times0.97}{0.1\times(1-\pi_1)\times0.01}}{\frac{0.9\times\pi_0\times0.97}{0.9\times(1-\pi_0)\times0.01}} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

- source: “Statistical models in Epidemiology”, by Clayton and Hills, page 156.

Why does $\widehat{OR}$ work? (2/2)

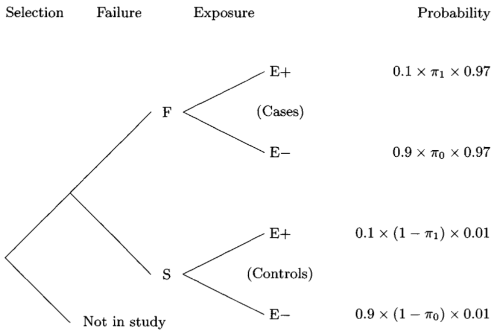

Selection
Failure
Exposure
Probability
F
(Cases)
E+
$0.1 \times \pi_1 \times 0.97$
E−
$0.9 \times \pi_0 \times 0.97$
S
(Controls)
E+
$0.1 \times (1-\pi_1) \times 0.01$
E−
$0.9 \times (1-\pi_0) \times 0.01$
Not in study

Fig. 16.2. The probability tree for the retrospective argument.

$$\widehat{OR} \approx \frac{\frac{0.1\times\pi_1\times0.97}{0.1\times(1-\pi_1)\times0.01}}{\frac{0.9\times\pi_0\times0.97}{0.9\times(1-\pi_0)\times0.01}} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

but

$$\widehat{RR} \approx \frac{\frac{0.1\times\pi_1\times0.97}{0.1\times\pi_1\times0.97\ +\ 0.1\times(1-\pi_1)\times0.01}}{\frac{0.9\times\pi_0\times0.97}{0.9\times\pi_0\times0.97\ +\ 0.9\times(1-\pi_0)\times0.01}} = \frac{\pi_1/(\pi_1\times0.97+(1-\pi_1)\times0.01)}{\pi_0/(\ \pi_0\times0.97+(1-\pi_0)\times0.01)} \neq \frac{\pi_1}{\pi_0}$$

- source: “Statistical models in Epidemiology”, by Clayton and Hills, page 156.

Outline/Intended Learning Outcomes (ILOs)

Preliminaries
- ILO: calculate 95% CIs for population proportions
- ILO: distinguish between exact and approximate (asymptotic) 95% CIs

Group comparison
- ILO: to define a suitable association measure and compute its 95% CI
- ILO: to (correctly) use the χ^2 test and Fisher's test

Sample size and power calculation
- ILO: to identify why and how to make power and sample size calculations
- ILO: to analyse their strengths and limitations

Confounding
- ILO: to exemplify confounding and its potential to be misleading
- ILO: to name two commonly used remedies

Cohort vs case-control study
- ILO: to differentiate the cohort and case-control designs
- ILO: to restate which association measure(s) can be used for each design

Screening: jargon
- ILO: to recognize some jargon

Paired binary data (if time allows)
- ILO: to exemplify paired binary data
- ILO: to calculate appropriate 95%-CI and p-values

Medical test / screening: jargon

Y: Outcome (disease status) E.g. prostate cancer

X: Test result (biomarker). E.g. $X = \begin{cases} 1 & \text{positive if PSA} > 4.0\,\text{ng/mL} \\ 0 & \text{negative if PSA} \leq 4.0\,\text{ng/mL} \end{cases}$

	$Y = 1$	$Y = 0$
$X = 1$	True positive	False positive
$X = 0$	False negative	True negative

- True positive rate (sensitivity): $P(X = 1 \mid Y = 1)$
- True negative rate (specificity): $P(X = 0 \mid Y = 0)$
- False positive rate (1-specificity): $P(X = 1 \mid Y = 0)$
- The positive predictive value: $P(Y = 1 \mid X = 1)$
- The negative predictive value: $P(Y = 0 \mid X = 0)$

Outline/Intended Learning Outcomes (ILOs)

Preliminaries
- ILO: calculate 95% CIs for population proportions
- ILO: distinguish between exact and approximate (asymptotic) 95% CIs

Group comparison
- ILO: to define a suitable association measure and compute its 95% CI
- ILO: to (correctly) use the χ^2 test and Fisher's test

Sample size and power calculation
- ILO: to identify why and how to make power and sample size calculations
- ILO: to analyse their strengths and limitations

Confounding
- ILO: to exemplify confounding and its potential to be misleading
- ILO: to name two commonly used remedies

Cohort vs case-control study
- ILO: to differentiate the cohort and case-control designs
- ILO: to restate which association measure(s) can be used for each design

Screening: jargon
- ILO: to recognize some jargon

Paired binary data (if time allows)
- ILO: to exemplify paired binary data
- ILO: to calculate appropriate 95%-CI and p-values

When do we typically meet paired binary data?

- **Comparison of diagnostic tests**
  - Example: compare sensitivity (i.e. True Positive Rate) of two diagnostic tests based on either Method 1 (e.g. Blood culture) or Method 2 (e.g. PCR: Polymerase Chain Reaction) using the the same blood samples (i.e. same patients).
- **Crossover clinical trials**
  - Example: compare two sedatives, w.r.t. proportions of side effects (e.g. not waking when fire alarm rings), each drug is given to each patient one evening (two evenings separated by one week). The same patients receive the two drugs.

Why does pairing matter?

- **Comparison of diagnostic tests**
  - Example (cont'): blood samples of "heavily" infected patients are easier to test positive than those of "mildly" infected patients. Hence, if one test is positive, the chance that the second test is positive is higher than expected in average.
- **Crossover clinical trials**
  - Example (cont'): some people sleep better than others. Some will never wake no matter what. Others are bad sleepers and will always wake. Hence, if a subject wakes the first night, the chance that he/she wakes up the second night is higher than expected in average.

Take home message: we expect less variability between two observations from the same patient than between two observations from two different patients. Appropriate statistical analysis will recognize this smaller variability. Less variability implies less random variation, which further implies more certainty, that is, narrower 95% CI and smaller p-values (than if the pairing was "wrongly" ignored).

How are paired data often presented?

- **Comparison of diagnostic tests**[24]
  - Example (cont'):

		PCR-test	
		Negative	Positive
BC-test	Negative	1	19
	Positive	2	2

Remarks:

1. This 2 by 2 table shows the pairing (and the raw data).
2. If the sensitivity of the two diagnostic tests are equally good, we expect (approx.) the same counts in the "upper right" and "lower left" cells.

[24] Example from: Nguyen et al. "Performance of Candida real-time polymerase chain reaction, β-D-glucan assay, and blood cultures in the diagnosis of invasive candidiasis." Clinical infectious diseases 54.9 (2012): 1240-1248.

Which statistical method with paired binary data?

- For p-value computation, we often use a **McNemar's test**
- Modern software can compute an "exact" version of the McNemar's test.
- An exact confidence interval can be computed for each of the two compared specificities (as seen in the first slides of the lecture)[25]

[25] Large sample (i.e. "approximate") confidence intervals can be computed for the difference in proportions (not shown in this course), but no "exact" method exists.

Which R code and conclusions?

```
library(exact2x2)                      # load a useful package
tab <- rbind(c(1,19),c(2,2))           # 2 by 2 table
mcnemar.exact(tab)                     # exact McNemar test
binom.test(x=sum(tab[,2]),n=sum(tab)) # sensitivity for PCR-test (95%-CI)
binom.test(x=sum(tab[2,]),n=sum(tab)) # sensitivity for BC-test  (95%-CI)
```

Conclusions:
The sensitivity of the PCR test (88%, 95%-CI=[68,97]) was found significantly higher than that of the blood culture test (17%, 95%-CI=[5,37]) among patients with deep-seated candidiasis (p-value<0.001).