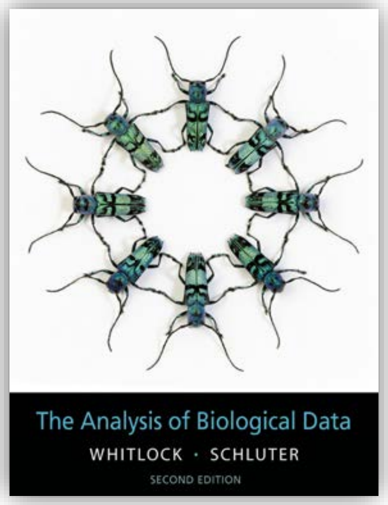
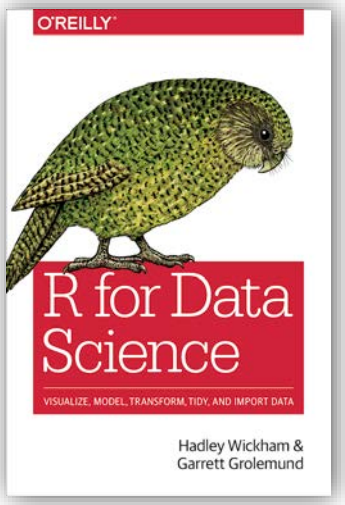
Data Science in Bioinformatics

Palle Villesen & Thomas Bataillon





Outline for week 06

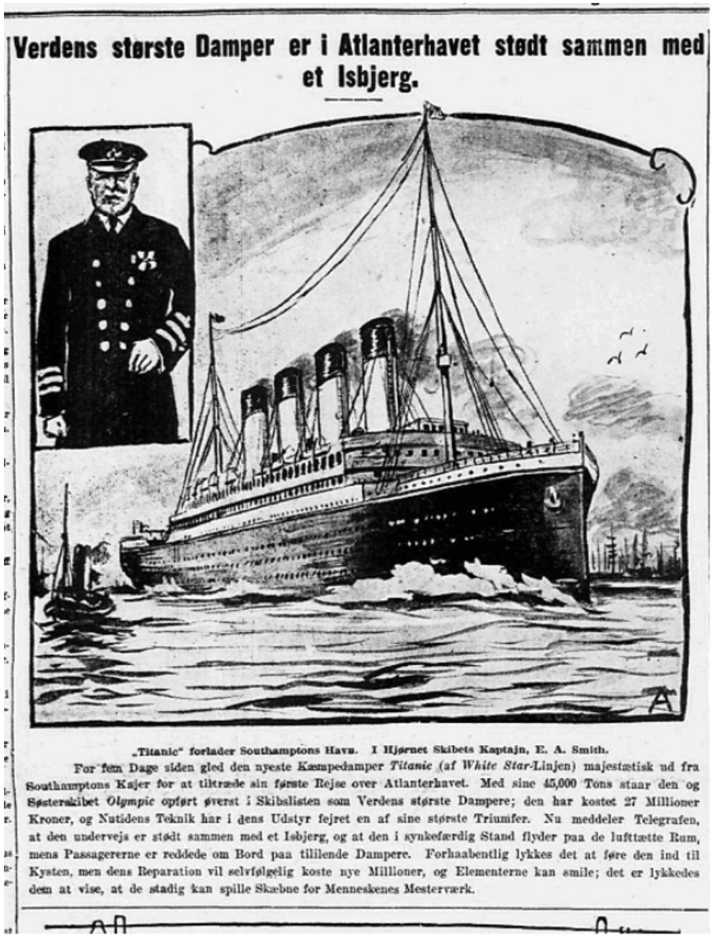
- Contingency tables
- Testing tables
- Working with big datasets
- The exercises

Contingency tables

- Basically tables on two categorical variables
- Simple examples
  - Headache/no headache in placebo/panodil group
  - Side effects/no side effects in placebo/vaccinated group
- Contingency analysis estimates and test for association between two or more categorical variables

Titanic 1912

- Male survivors = 367
- Female survivors = 344



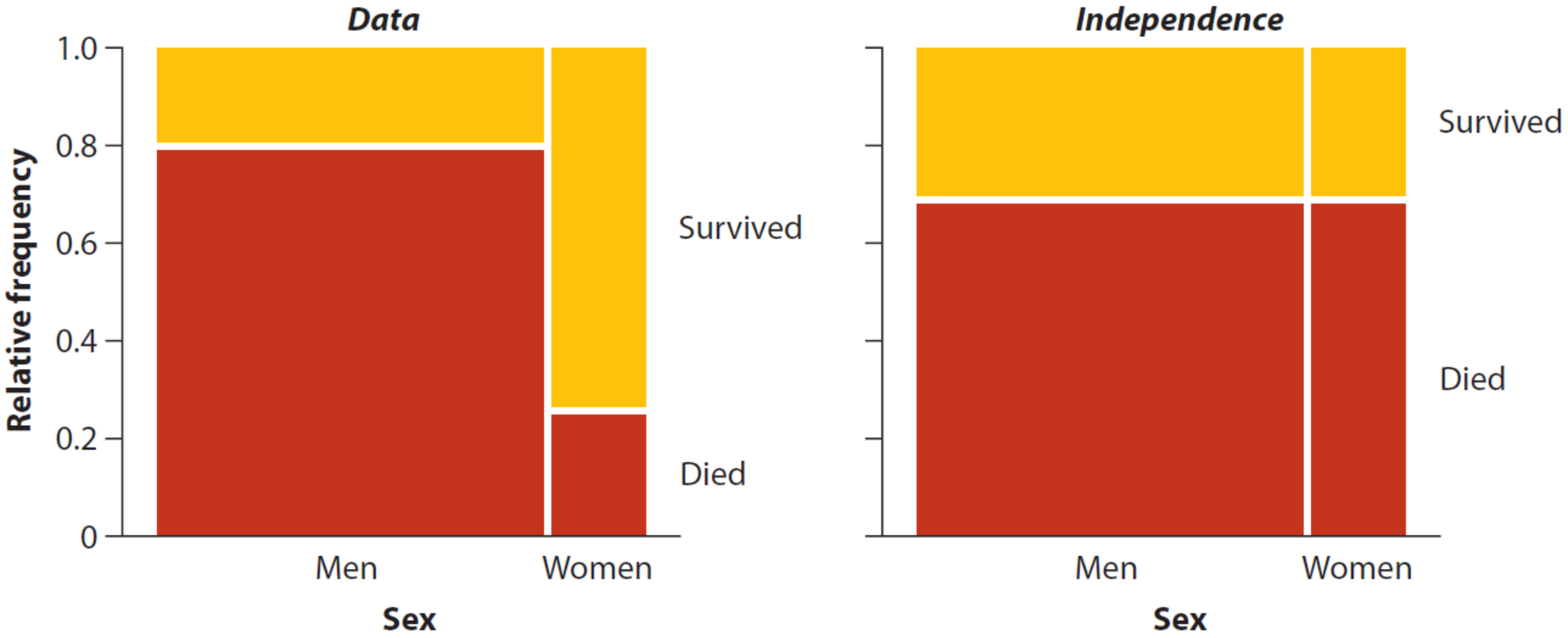
Verdens største Damper er i Atlanterhavet stødt sammen med et Isbjerg.

„Titanic" forlader Southamptons Havn. I Hjørnet Skibets Kaptajn, E. A. Smith.

For fem Dage siden gled den nyeste Kæmpedamper *Titanic* (af *White Star*-Linjen) majestætisk ud fra Southamptons Kajer for at tiltræde sin første Rejse over Atlanterhavet. Med sine 45,000 Tons staar den og Søsterskibet *Olympic* opført øverst i Skibslisten som Verdens største Dampere; den har kostet 27 Millioner Kroner, og Nutidens Teknik har i dens Udstyr fejret en af sine største Triumfer. Nu meddeler Telegrafen, at den undervejs er stødt sammen med et Isbjerg, og at den i synkefærdig Stand flyder paa de lufttætte Rum, mens Passagererne er reddede om Bord paa tililende Dampere. Forhaabentlig lykkes det at føre den ind til Kysten, men dens Reparation vil selvfølgelig koste nye Millioner, og Elementerne kan smile; det er lykkedes dem at vise, at de stadig kan spille Skæbne for Menneskenes Mesterværk.

https://www.kaggle.com/c/titanic

Titanic



```
> chisq.test(TitanicTable, correct=F)

	Pearson's Chi-squared test

data:  TitanicTable
X-squared = 456.87, df = 1, p-value < 2.2e-16
```

Odds

- $O = \frac{p}{1-p}$
- p = probability of ”success” in the group
  - Very often we call this ”risk of…”
- $\text{Odds ratio } = \frac{O_1}{O_2}$
- $\text{Relative risk } = \frac{p_1}{p_2}$

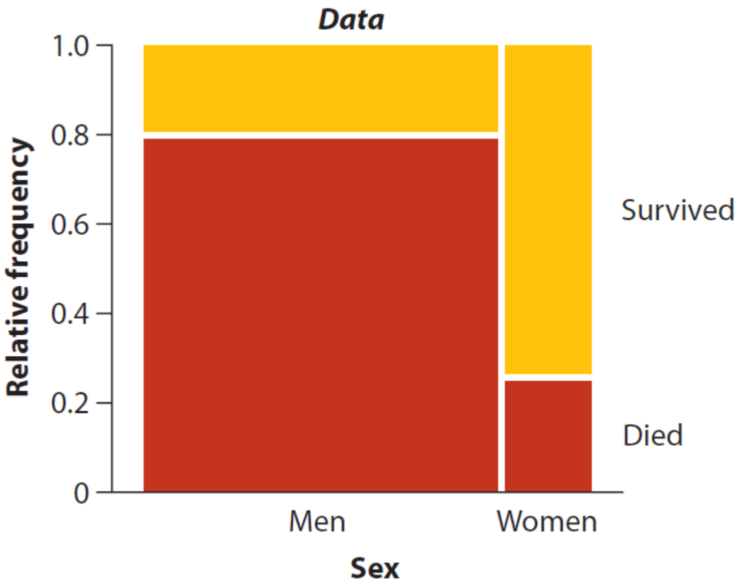
Titanic

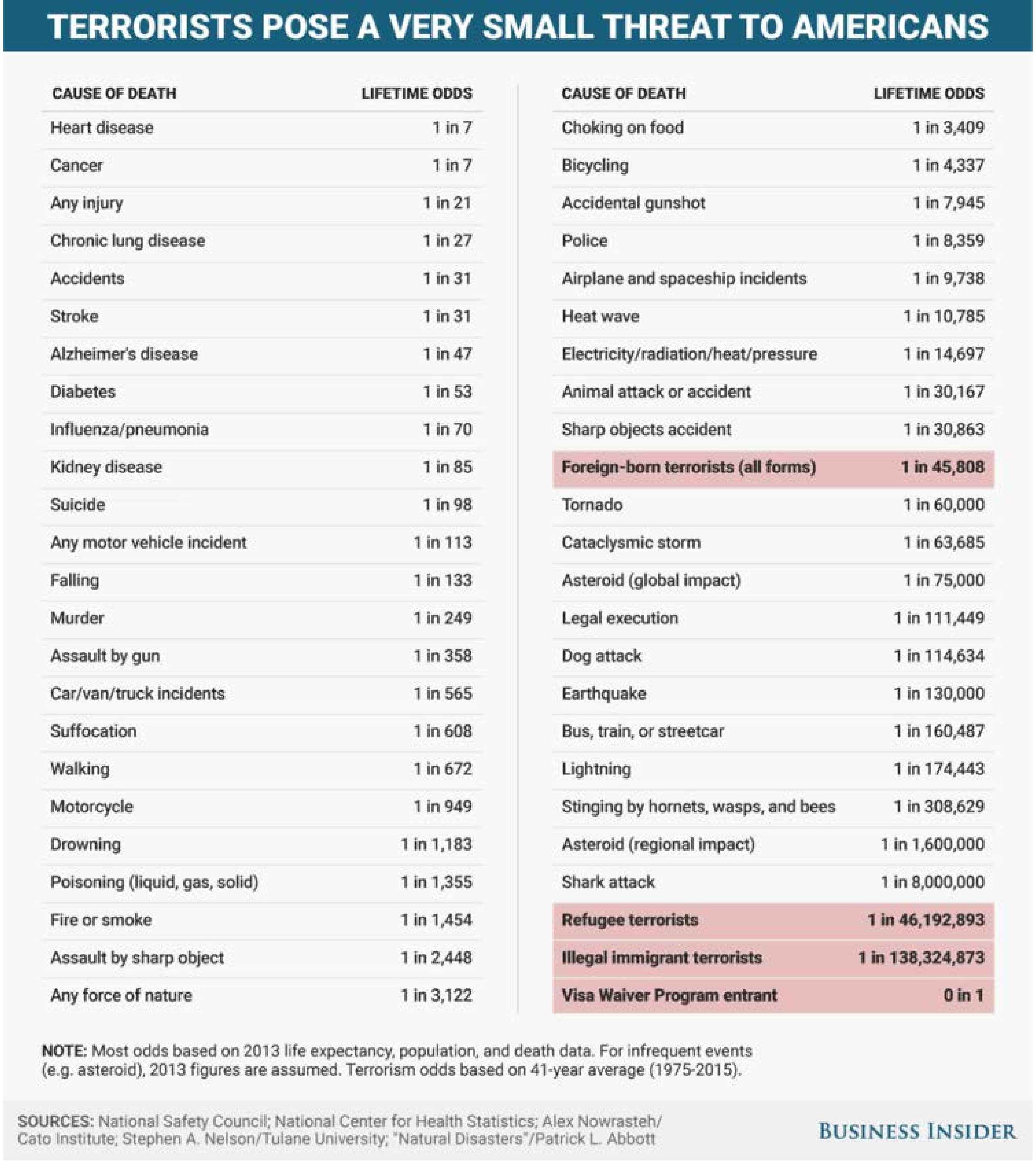
```
> TitanicTable
      Survived Died
Men        367 1364
Women      344  126

> p1 = 1364 / (1364+367)
> o1 = p1 / (1-p1)
> o1
[1] 3.716621
> p2 = 126 / (126+344)
> o2 = p2 / (1-p2)
> o2
[1] 0.3662791
> o1/o2
[1] 10.14697
```

Odds ratio

```
> p1/p2
[1] 2.939305
```

Relative risk





TERRORISTS POSE A VERY SMALL THREAT TO AMERICANS

CAUSE OF DEATH	LIFETIME ODDS	CAUSE OF DEATH	LIFETIME ODDS
Heart disease	1 in 7	Choking on food	1 in 3,409
Cancer	1 in 7	Bicycling	1 in 4,337
Any injury	1 in 21	Accidental gunshot	1 in 7,945
Chronic lung disease	1 in 27	Police	1 in 8,359
Accidents	1 in 31	Airplane and spaceship incidents	1 in 9,738
Stroke	1 in 31	Heat wave	1 in 10,785
Alzheimer's disease	1 in 47	Electricity/radiation/heat/pressure	1 in 14,697
Diabetes	1 in 53	Animal attack or accident	1 in 30,167
Influenza/pneumonia	1 in 70	Sharp objects accident	1 in 30,863
Kidney disease	1 in 85	**Foreign-born terrorists (all forms)**	**1 in 45,808**
Suicide	1 in 98	Tornado	1 in 60,000
Any motor vehicle incident	1 in 113	Cataclysmic storm	1 in 63,685
Falling	1 in 133	Asteroid (global impact)	1 in 75,000
Murder	1 in 249	Legal execution	1 in 111,449
Assault by gun	1 in 358	Dog attack	1 in 114,634
Car/van/truck incidents	1 in 565	Earthquake	1 in 130,000
Suffocation	1 in 608	Bus, train, or streetcar	1 in 160,487
Walking	1 in 672	Lightning	1 in 174,443
Motorcycle	1 in 949	Stinging by hornets, wasps, and bees	1 in 308,629
Drowning	1 in 1,183	Asteroid (regional impact)	1 in 1,600,000
Poisoning (liquid, gas, solid)	1 in 1,355	Shark attack	1 in 8,000,000
Fire or smoke	1 in 1,454	**Refugee terrorists**	**1 in 46,192,893**
Assault by sharp object	1 in 2,448	**Illegal immigrant terrorists**	**1 in 138,324,873**
Any force of nature	1 in 3,122	**Visa Waiver Program entrant**	**0 in 1**

NOTE: Most odds based on 2013 life expectancy, population, and death data. For infrequent events (e.g. asteroid), 2013 figures are assumed. Terrorism odds based on 41-year average (1975-2015).

SOURCES: National Safety Council; National Center for Health Statistics; Alex Nowrasteh/Cato Institute; Stephen A. Nelson/Tulane University; "Natural Disasters"/Patrick L. Abbott

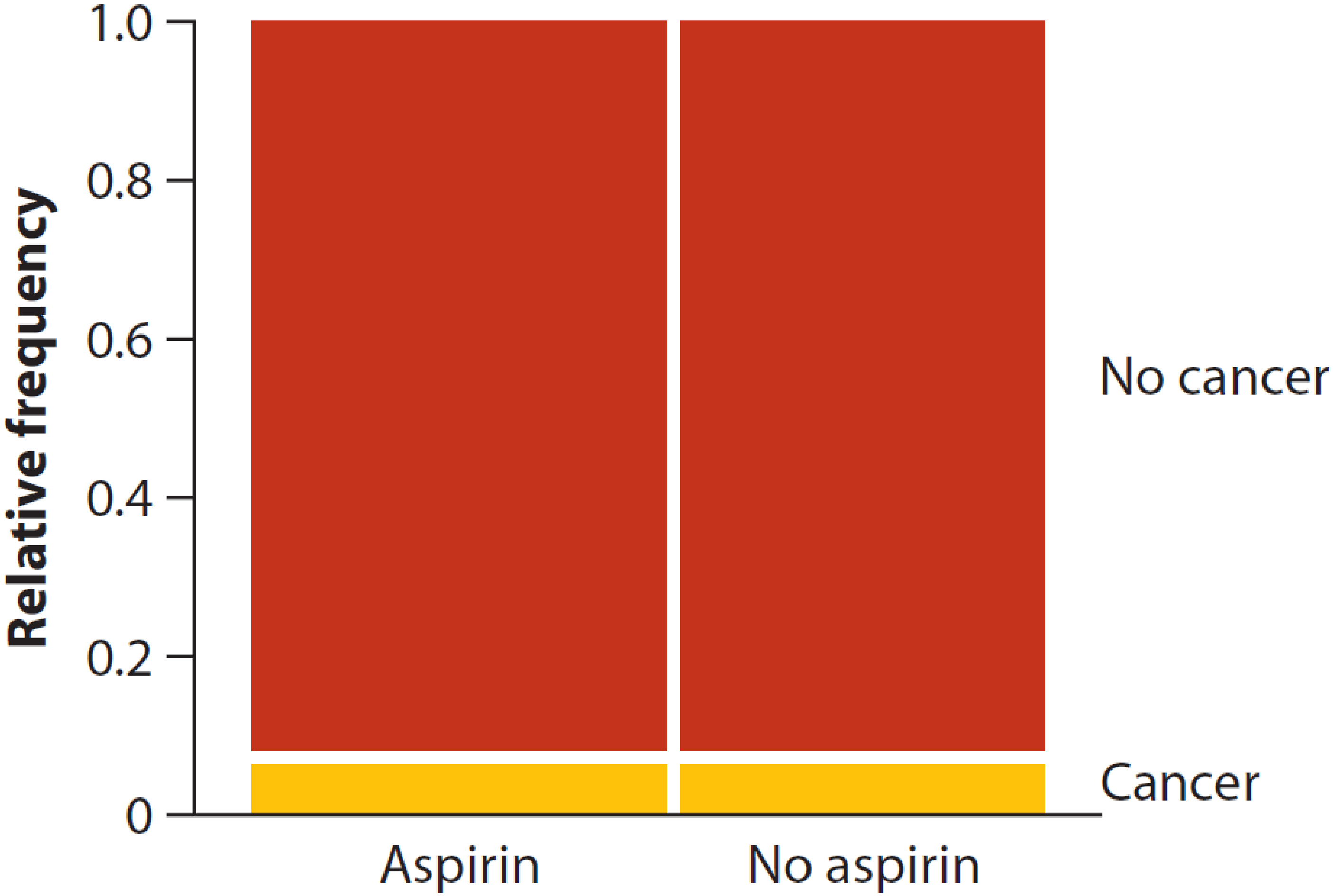
BUSINESS INSIDER

Aspirin and cancer

Zum Einnehmen mit reichlich Flüssigkeit.
Anwendungsgebiete: Leichte bis mäßig starke Schmerzen, Fieber.
Arzneimittel für Kinder unzugänglich aufbewahren!
Apothekenpflichtig. Nicht über 30 °C lagern.
BAYER
ASPIRIN®
Tablette mit 500 mg Acetylsalicylsäure
20 Tabletten
ASPIRIN

	Aspirin	Placebo
Cancer	1438	1427
No cancer	18496	18515

Aspirin and cancer (observed)



Odds of cancer on aspirin

- $O = \frac{p}{1-p}$
- p(cancer on aspirin) $= \frac{1438}{1438+18496} = 0.0721$
- $1-\mathrm{p} = 0.9279$
- $O_1 = \frac{p}{1-p} = \frac{0.0721}{0.9279} = 0.0777$

	Aspirin	Placebo
Cancer	1438	1427
No cancer	18496	18515

Odds of cancer on placebo

- $O = \frac{p}{1-p}$
- p(cancer on placebo) $= \frac{1427}{1427+18515} = 0.0716$
- $1-\mathrm{p} = 0.9285$
- $O_2 = \frac{p}{1-p} = \frac{0.0716}{0.9285} = 0.0771$

	Aspirin	Placebo
Cancer	1438	1427
No cancer	18496	18515

Odds ratio

- The ratio of the odds (the aspirin vs. the placebo group)
- $OR = \frac{O_1}{O_2}$
- $O_1 = \frac{p}{1-p} = \frac{0.0721}{0.9279} = 0.0777$
- $O_2 = \frac{p}{1-p} = \frac{0.0716}{0.9285} = 0.0771$
- Odds ratio = 1.008

Standard error of OR

```
OR = (a*d) /( b*c)
OR = (1438*18515)/(1427*18496) = 1.009
ln(OR) = ln(1.009) = 0.00896
SE[ln(OR)] = sqrt(1/a + 1/b + 1/c + 1/d)
           = 0.03878
```

	Aspirin	Placebo
Cancer	a = 1438	b = 1427
No cancer	c = 18496	d = 18515

Confidence interval of OR

$$\ln(OR)-1.96*SE < \ln(OR) < \ln(OR)+1.96*SE$$

$$-0.067 < \ln(OR)) < 0.085$$

$$\exp(-0.067) < \exp(\ln(OR))< \exp(0.085)$$

$$0.93 < OR < 1.09$$

	Aspirin	Placebo
Cancer	a = 1438	b = 1427
No cancer	c = 18496	d = 18515

Case-control studies

- Two groups of samples
  - 1000 schizophrenia patiens
  - 1000 controls
- The group size is chosen by us – not by the frequency in the poulation
  - So one of the groups is LARGER than in real life

	Likes coca cola	Hates coca cola
Schizophrenia	800	200
Normal	700	300

Relative risk

- RR = p1 / p2
- Only possible when p1 and p2 are unbiased estimates
- NOT possible for this dataset.

	Likes coca cola	Hates coca cola
Schizophrenia	800	200
Normal	700	300

5 minutes

- What is the response variable (outcome) in the example
- What is the explanatory variable in the example?
- Assume case/control is response
  - What is the problem with estimating p1?
- What is the OR?

	Likes coca cola	Hates coca cola
Schizophrenia	800	200
Normal	700	300

Relative risk

- RR = p1 / p2
- Only possible when you know p1
- Coca cola drinker and schizophrenia outcome
- p1 = risk of schizo if you like coca cola
  - We do not know this…
- But we can calculate OR = 24/14 = 1.71

	Likes coca cola	Hates coca cola
Schizophrenia	800	200
Normal	700	300

Relative risk

- RR = p1 / p2
- Calculate RR for the following prevalences of schizophrenia
  - 1:1 (so 1000 schizophrenia and 1000 normal)
  - 1:10 (so 100 schizophrenia and 1000 normal)
  - 1:100 (so 10 schizophrenia and 1000 normal)
  - 1:1000 (so 1 schizophrenia and 1000 normal)

	Likes coca cola	Hates coca cola
Schizophrenia	800	200
Normal	700	300

Relative risk and OR

- P1 is biased (schizo:healthy)

```
> p1=800/1500
> p2=200/500
> p1/p2
[1] 1.333333
>
> p1=80/780
> p2=20/320
> p1/p2
[1] 1.641026
>
> p1=8/708
> p2=2/302
> p1/p2
[1] 1.706215
>
> p1=.8/700.8
> p2=.2/300.2
> p1/p2
[1] 1.71347
```

	Likes coca cola	Hates coca cola
Schizophrenia	800	200
Normal	700	300

χ^2 contingency test

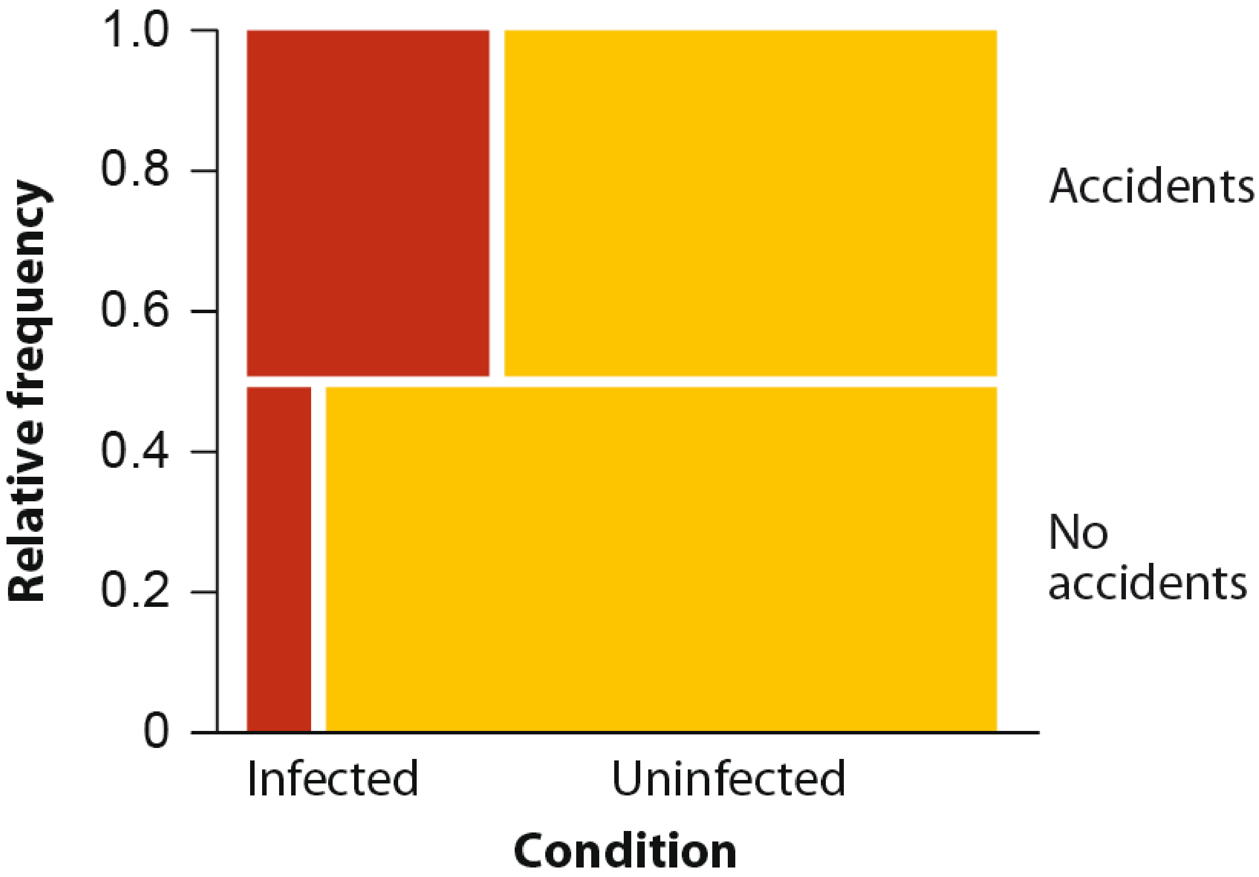
- One assumption that is really important
  - A common rule is 5 or more in all cells of a 2-by-2 table, and 5 or more in 80% of cells in larger tables, but no cells with zero expected count.
- When this assumption is not met
  - Fishers exact test

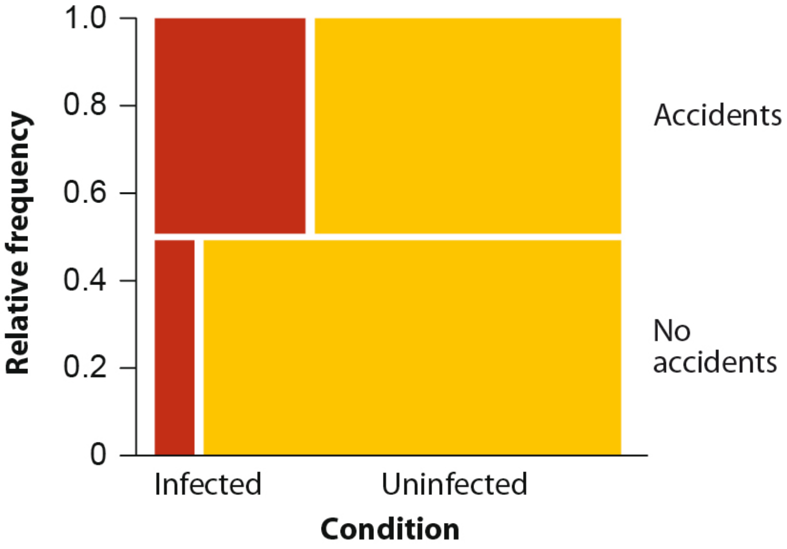
- R: chisq.test(table(), correct=F)
- It is basically a goodness of fit test.

Cats are evil

	infected	uninfected
accidents	61	124
no accidents	16	169

Toxoplasma gondii OR=5.20





```
> chisq.test(x = toxTable)

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

data:  toxTable
X-squared = 31.75, df = 1, p-value = 1.753e-08

>
```

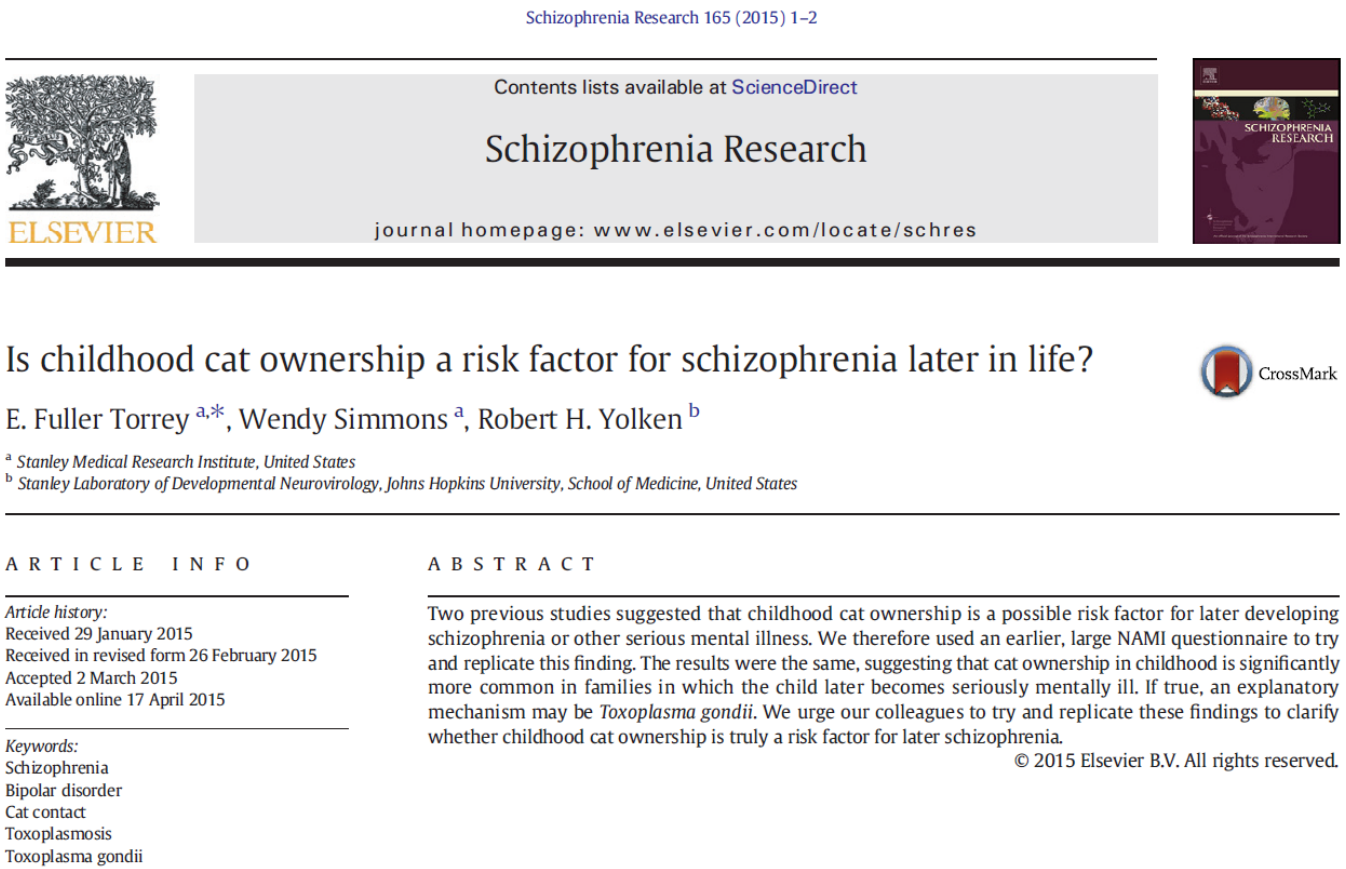
Example 9.4

- You will do this during the exercises

Table 1
Cat ownership in NAMI families and controls.

		Cases	Controls	
1992 questionnaire	Cat in house, birth to age 10	84/165 (50.9%)	65/165 (39.4%)	$p = .03$; OR = 1.60 (1.00–2.53)
1997 survey	Cat ownership, birth to age 13	136/262 (51.9%)	220/522 (42.1%)	$p = .01$; OR = 1.48 (1.09–2.02)
1982 questionnaire	Cat ownership, birth to age 13	1075/2125 (50.6%)	2065/4847 (42.6%)	$p \leq .0001$; OR = 1.38 (1.25–1.53)

p values are derived from chi square, 2 tailed; ORs shown as mean (95% CI).



Schizophrenia Research 165 (2015) 1–2

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Is childhood cat ownership a risk factor for schizophrenia later in life?

E. Fuller Torrey [a,*], Wendy Simmons [a], Robert H. Yolken [b]

[a] *Stanley Medical Research Institute, United States*
[b] *Stanley Laboratory of Developmental Neurovirology, Johns Hopkins University, School of Medicine, United States*

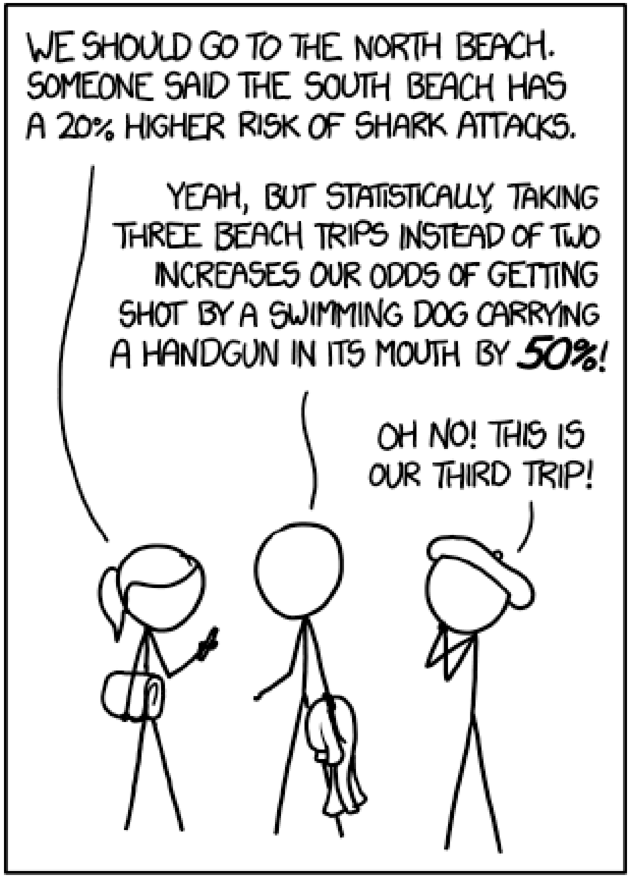
ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form 26 February 2015
Accepted 2 March 2015
Available online 17 April 2015

Keywords:
Schizophrenia
Bipolar disorder
Cat contact
Toxoplasmosis
Toxoplasma gondii

ABSTRACT

Two previous studies suggested that childhood cat ownership is a possible risk factor for later developing schizophrenia or other serious mental illness. We therefore used an earlier, large NAMI questionnaire to try and replicate this finding. The results were the same, suggesting that cat ownership in childhood is significantly more common in families in which the child later becomes seriously mentally ill. If true, an explanatory mechanism may be *Toxoplasma gondii*. We urge our colleagues to try and replicate these findings to clarify whether childhood cat ownership is truly a risk factor for later schizophrenia.

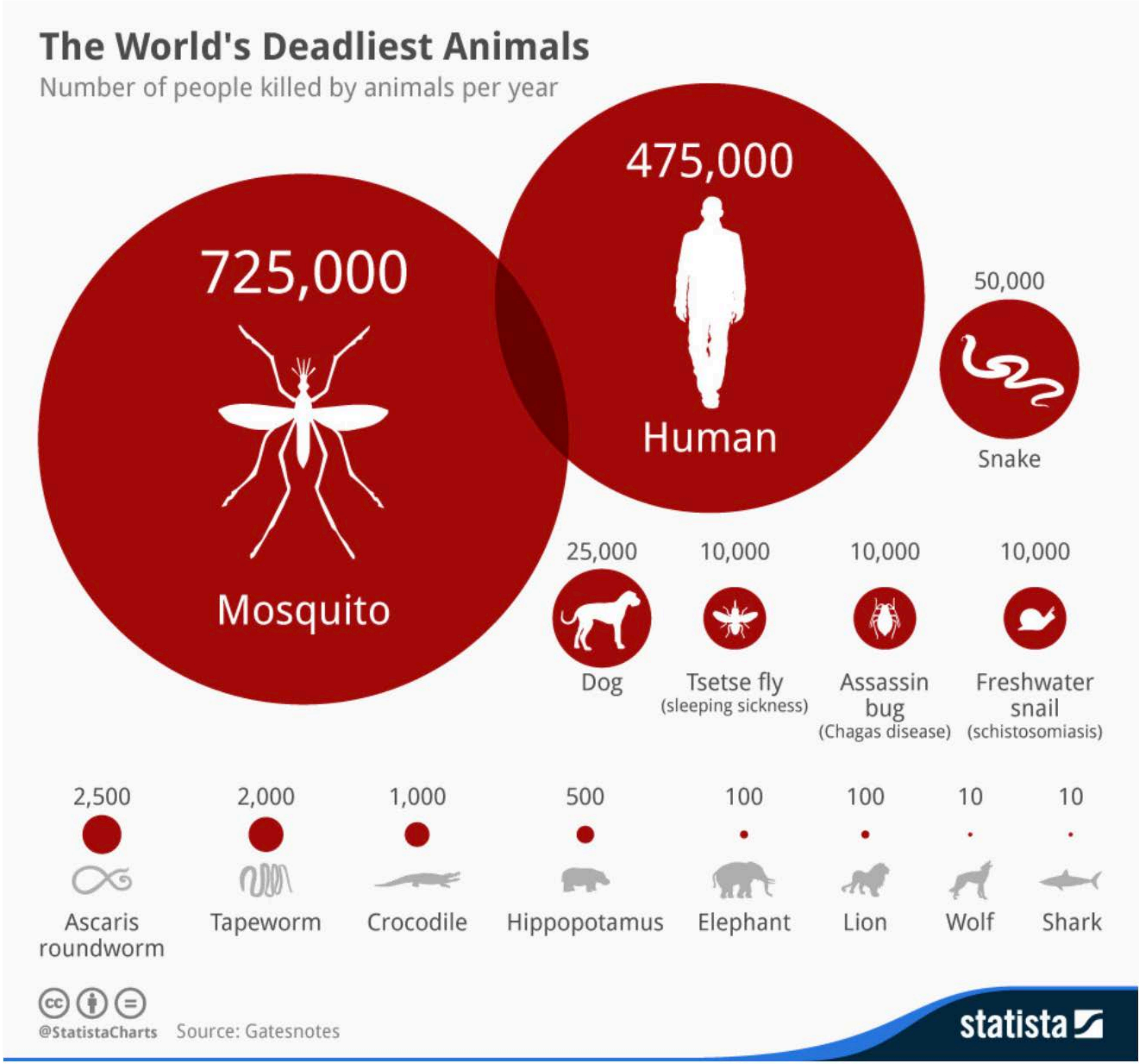

WE SHOULD GO TO THE NORTH BEACH. SOMEONE SAID THE SOUTH BEACH HAS A 20% HIGHER RISK OF SHARK ATTACKS.
YEAH, BUT STATISTICALLY, TAKING THREE BEACH TRIPS INSTEAD OF TWO INCREASES OUR ODDS OF GETTING SHOT BY A SWIMMING DOG CARRYING A HANDGUN IN ITS MOUTH BY *50%!*
OH NO! THIS IS OUR THIRD TRIP!

REMINDER: A 50% INCREASE IN A TINY RISK IS ***STILL TINY.***

Breast cancer and noise

- Up to 28% higher risk pr. 10 dB for estrogen receptor negative BC
- 30000 women
- 1219 with BC
- 203 with ER- (estrogen receptor negative)
- So increase in risk is from ~0.0067 to ~0.0086

- “We found no overall association between residential road traffic or railway noise and breast cancer risk.”



The World's Deadliest Animals

Number of people killed by animals per year

@StatistaCharts Source: Gatesnotes

Working with larger datasets

- Walkthrough of one rmarkdown html file
  - The sqlite is for future use (if any)
- This weeks exercises.

Too much time?

- Read r help
  - http://whitlockschluter.zoology.ubc.ca/r-code/rcode09
- Get data on toxoplasma
  - http://whitlockschluter.zoology.ubc.ca/data/chapter09
- Discuss
  - Is this tidy data or not?
  - Calculate Odd ratio
  - Test the association