Measures of Association

Applied Veterinary Epidemiology for Ghana

EPIDEMIOLOGY, ECONOMICS AND RISK ASSESSMENT (EERA)

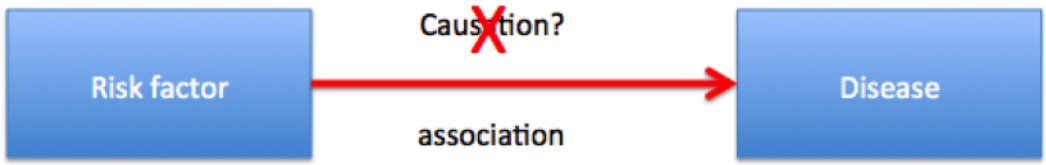
Part 3: Measures of Association

3.1 Study design

Observational study designs

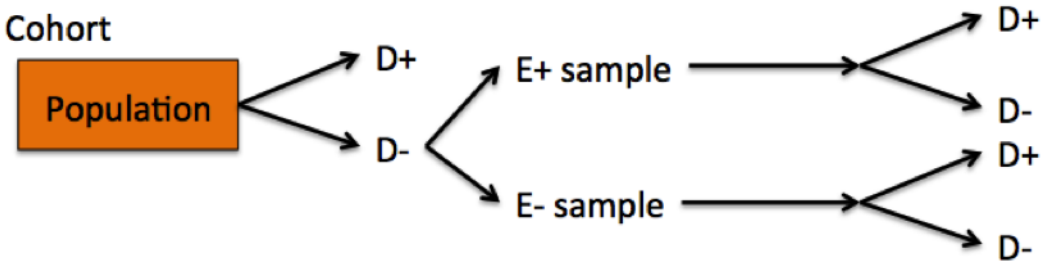
Observational studies look at how risk factors are related to a specific outcome, such as a disease, without changing which subjects are exposed or unexposed to the risk factors. This is in contrast to experimental study designs, where the subjects are assigned to groups and then exposed to a risk factor eg. some subjects get treatment A, others get treatment B. There are three main study designs for observational studies: cohort studies, case-control studies, and cross sectional studies. The type of study used is dependent on the research question and the resources available for the study.

But a key concept here is that these statistical tools tell you about statistical ASSOCIATIONS. This is not the same as CAUSATION!.



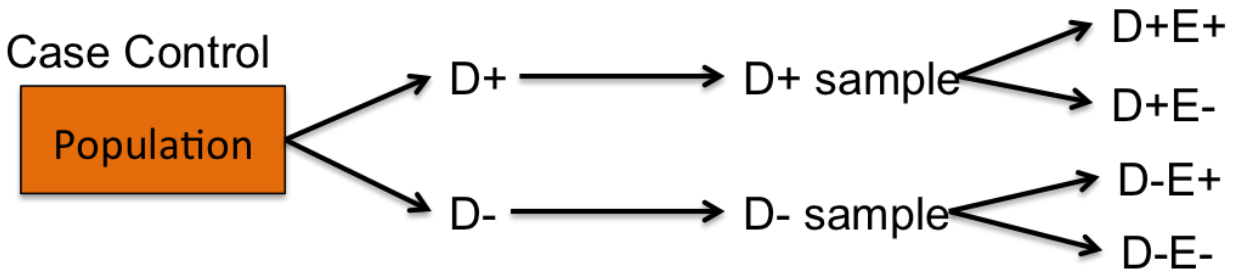
Cohort studies

- the selection of the study population is based on exposure or through identifying a population and watching to see who gets exposed and who develops the disease
- longitudinal study or can shorten if data on exposure from historical records available
- very good for looking at the timing of exposures and disease (important for causation)
- often very expensive and longer duration than most funders prepared to pay for
- the lack of randomisation of individuals to exposures is where epidemiology diverges from experimental studies or clinical trials
- can estimate relative risk or odds ratios



Case control studies

- differs fundamentally from cohort study as you start with the individuals with the disease (*cases*) and then look for controls
- you then determine the proportion of cases (and controls) who have the exposure of interest
- NB. you can not estimate the prevalence of disease from a case control study as you have determined the proportions of cases and controls
- selection of controls is the main challenge with case-control studies
- controls should represent the rate of exposure of interest in the general population and also have the potential to become cases
- sources of controls:
  1. hospital patients (e.g. trauma patients or specific groups - often unrepresentative of the reference population)
  2. non-hospital patients (e.g. random digit dialling, neighbourhood or best friend)
- matching controls to cases
  - e.g. by age group, sex etc.
  - may be difficult to find controls if match on too many characteristics
  - if you match on a particular characteristic you can not study it
  - if you match you will control confounding for the things you match on but CAN introduce confounding on other characteristics unknowingly
  - can ONLY estimate odds ratios



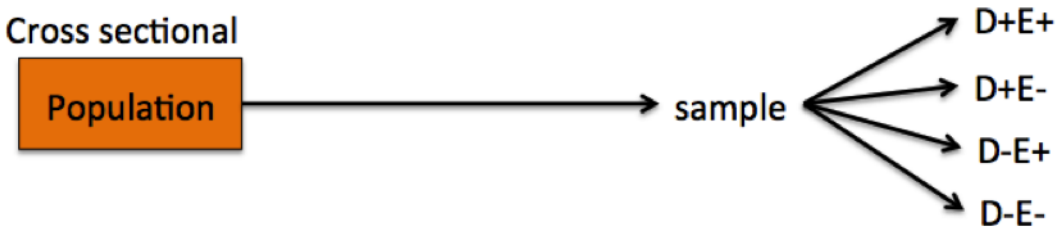
NB. you will see in the literature that people refer to these as prospective and retrospective studies. Do not use these terms as they are inaccurate and misleading and confusing.

A hybrid of the cohort and case control study is the nested case control study where a cohort study is run and then cases identified and controls drawn from the cohort. This has several advantages:

- data collected from the start and samples collected from the start and stored
- more economic as you do not test all samples in cohort

Cross sectional studies

- very commonly used
- both disease status and exposure determined at the same time
- the cases of disease we see are prevalent cases of the disease (WARNING: we may be missing those that died acutely!!)
- because of the design it is generally not possible to determine the ordering of events with great certainty and so causality can be very difficult to interpret
- can estimate relative risk or odds ratios



Randomised control trial

- this is an experimental study design and not an observational design.

3.2 Estimating Risk and Associations

The overall risk of a disease is also known as the absolute risk of the disease. This is the proportion of the population at risk that develop the disease.

However, in general we are interested in estimating the association between an exposure and disease as we are interested in knowing what things increase (or decrease) our risk of developing a disease and if we implement some intervention how many fewer cases we are likely to see.

To measure this association we will look at the following: relative risk, incidence rate ratios, odds ratios (commonly used), and prevalence ratios.

Relative risk

	D+'ve	**D-'ve**	total
Exp +'ve	a	b	a+b (n1)
Exp -'ve	c	d	c+d (n2)

If we start with a cohort study we can look at the risk of getting the disease if you have the exposure $\frac{a}{a+b}$ and compare this with the risk if you did not have the exposure $\frac{c}{c+d}$. An important point to realise in the design is that you have two groups sampled; exposed and unexposed, and you see how many develop disease, so the proportion that develop disease in each group is an estimate of the proportion in the population.

The relative risk is then the ratio of the risk in the exposed group compared to the unexposed group as follows:

$$relative\,risk = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

If the relative risk is 1, then there is no evidence of an association between the exposure and the disease. If the relative risk is greater than 1, then there is a greater risk of disease in the exposed group than in the unexposed group. If the relative risk is less than 1, then exposure may be associated with a decreased risk of disease.

Import the cattle data from earlier and have a look at the relative risk of being positive for Lepto if you are male (1) compared to being female (2).

```
library(tidyverse)
library(here)
library(epiDisplay)
library(epiR)
```

```
#Get the data back into RStudio with the code from before and clean it up if needed
dat <- read_csv(here("data", "cattle_data.csv"))
```

```
## Rows: 199 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): DamID
## dbl (14): Lepto, BrucellaLFA, BrucellaRBG, GirthDam, CSDam, SL, East, North,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Have a look at the data
head(dat)
```

```
## # A tibble: 6 x 15
##   DamID    Lepto BrucellaLFA BrucellaRBG GirthDam CSDam    SL  East North Elevat
##   <chr>    <dbl>       <dbl>       <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 DM01012~     1           0           0     140    7       1  34.5 0.637   1446
## 2 DM01012~     0           0           0     144    7       1  34.5 0.627   1360
## 3 DM01012~     0           0           0     152.   6.5     1  34.5 0.634   1349
## 4 DM01012~     0           0           0     147    5       1  34.5 0.633   1379
## 5 DM01012~     0           0           0     142.   6       1  34.5 0.623   1344
## 6 DM01012~     0           0           0     142.   7       1  34.5 0.628   1363
## # i 5 more variables: CalfSex <dbl>, DAge <dbl>, DCalving <dbl>, TDFarm <dbl>,
## #   DAge2 <dbl>
```

```
# Data cleaning (as in previous session)
dat <- dat %>%
      mutate(North = case_when(North == 0.060574 ~ 0.60574,
                               TRUE ~ North),
             East = case_when(East == 3.446714 ~ 34.46714,
                              TRUE ~ East)) %>%
      mutate(DAge = na_if(DAge, -888),
             DCalving = na_if(DCalving, -888),
             TDFarm = na_if(TDFarm, -888),
             CalfSex = case_when(as.factor(CalfSex) == 1 ~ "Male",
                                 as.factor(CalfSex) == 2 ~ "Female"))
```

We will try the *epi.2by2* function from the *epiR* package.

```
cc(dat$CalfSex, dat$Lepto)
```

```
##
##            dat$Lepto
## dat$CalfSex   0   1 Total
##      Female  94   4    98
##      Male    93   8   101
##      Total  187  12   199
##
## OR =  2.02
## 95% CI =  0.59, 6.94
## Chi-squared = 1.29, 1 d.f., P value = 0.255
## Fisher's exact test (2-sided) P value = 0.373
```

```
tab <- table(dat$Lepto, dat$CalfSex)
# exp_case, exp_noncase, nonexp_case, nonexp_noncase
x <- c(tab[2,2], tab[1,2], tab[2,1], tab[1,1])
# you will need to check this order for any given table to make sure it is the comparison you want

epi.2by2(x, method = "cohort.count")
```

```
##              Outcome +    Outcome -      Total                 Inc risk *
## Exposed +            8           93        101       7.92 (3.48 to 15.01)
## Exposed -            4           94         98       4.08 (1.12 to 10.12)
## Total               12          187        199       6.03 (3.15 to 10.30)
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 1.94 (0.60, 6.24)
## Inc odds ratio                                 2.02 (0.59, 6.94)
## Attrib risk in the exposed *                   3.84 (-2.72, 10.40)
## Attrib fraction in the exposed (%)            48.47 (-65.63, 83.97)
## Attrib risk in the population *                1.95 (-3.18, 7.08)
## Attrib fraction in the population (%)         32.31 (-47.61, 68.96)
## -------------------------------------------------------------------
## Uncorrected chi2 test that OR = 1: chi2(1) = 1.294 Pr>chi2 = 0.255
## Fisher exact test that OR = 1: Pr>chi2 = 0.373
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 population units
```

Have a look at the output. You can see that we have estimated the risk ratio, odds ratio and attributable risks. Is it important to note that the output for the *epi.2by2* function will differ depending on whether you define method as a 'cohort.count', 'cohort.time', 'case.control', or 'cross.sectional'.

Looking at incidence If you have a full longitudinal dataset where you have observed the animals for a number of years we can include the time aspect formally (and thus compare incidence rates) and we need to use a slightly different analysis. Here we look at the total time that the individuals were at risk for, rather than just the total number of individuals at risk.

	D+'ve	**time at risk**	
Exp +'ve	a	t1	(n1)
Exp -'ve	c	t2	(n2)

```r
tab <- dat %>%
  group_by(CalfSex) %>%
  summarise(
    Dis = sum(Lepto),
    Time = sum(DAge2)
  )
tab
```

```
## # A tibble: 2 x 3
##   CalfSex   Dis  Time
##   <chr>   <dbl> <dbl>
## 1 Female      4  713
## 2 Male        8  722.
```

```r
x <- as.numeric(c(tab[2,2], tab[2,3], tab[1,2], tab[1,3]))
# you will need to check this order for any given table to make sure it is the comparison you want

epi.2by2(x, method="cohort.time")
```

```
##              Outcome +    Time at risk                 Inc rate *
## Exposed +            8           721.5        1.11 (0.48 to 2.18)
## Exposed -            4             713        0.56 (0.15 to 1.44)
## Total               12          1434.5        0.84 (0.43 to 1.46)
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc rate ratio                                 1.98 (0.53, 8.97)
## Attrib rate in the exposed *                   0.55 (-0.40, 1.49)
## Attrib fraction in the exposed (%)            49.40 (-88.86, 88.85)
## Attrib rate in the population *                0.28 (-0.45, 1.00)
## Attrib fraction in the population (%)         32.94 (1.70, 64.64)
## -------------------------------------------------------------------
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 units of population time at risk
```

```r
#Or just put in the numbers manually.
epi.2by2(c(8, 722, 4, 713), method="cohort.time")
```

```
##              Outcome +    Time at risk                 Inc rate *
## Exposed +            8             722        1.11 (0.48 to 2.18)
## Exposed -            4             713        0.56 (0.15 to 1.44)
## Total               12            1435        0.84 (0.43 to 1.46)
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc rate ratio                                 1.98 (0.53, 8.96)
## Attrib rate in the exposed *                   0.55 (-0.40, 1.49)
## Attrib fraction in the exposed (%)            49.37 (-88.99, 88.84)
## Attrib rate in the population *                0.28 (-0.45, 1.00)
## Attrib fraction in the population (%)         32.91 (1.67, 64.62)
## -------------------------------------------------------------------
```

```
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 units of population time at risk
```

One more example on time at risk:

	mastitis+'ve	time at risk	
not predipped (Exp+'ve)	a = 18	b = 250	(n1)
predipped (Exp-'ve)	c = 8	d = 236	(n2)

```
epi.2by2(c(18,250,8,236), method="cohort.time")
```

```
##              Outcome +    Time at risk                 Inc rate *
## Exposed +           18             250       7.20 (4.27 to 11.38)
## Exposed -            8             236        3.39 (1.46 to 6.68)
## Total               26             486        5.35 (3.49 to 7.84)
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc rate ratio                                 2.12 (0.88, 5.65)
## Attrib rate in the exposed *                   3.81 (-0.26, 7.88)
## Attrib fraction in the exposed (%)            52.92 (-13.80, 82.29)
## Attrib rate in the population *                1.96 (-1.16, 5.08)
## Attrib fraction in the population (%)         36.64 (14.79, 58.12)
## -------------------------------------------------------------------
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 units of population time at risk
```

Odds ratios

You can calculate an odds ratio for all 3 types of observational studies but for a case control study it is the only measure you can estimate because you fixed the numbers of cases and controls and therefore can not estimate the risk of disease in each exposure group anymore. However, you can compare the relative exposures in case and controls.

	D+'ve	D-'ve
Exp +'ve	a	b
Exp -'ve	c	d
—	n1	n2

The formula for estimating the odds ratio looks like this. N.B. that this avoids adding across the columns which would be meaningless.

$$Odds\ ratio = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \times d}{b \times c}$$

If the odds ratio is 1, then there is no evidence of an association between the exposure and the disease. If the odds ratio is greater than 1, then there is a greater odds of disease in the exposed group than in the

unexposed group. If the odds ratio is less than 1, then exposure may be associated with a decreased odds of disease, meaning that the factor may have a protective effect.

The odds ratio can be very similar to the RR in situations when there is a rare disease.

```
tab <- table(dat$Lepto, dat$CalfSex)
# exp_case, exp_noncase, nonexp_case, nonexp_noncase
x <- c(tab[2,2], tab[1,2], tab[2,1], tab[1,1])

epi.2by2(x, method = "case.control")
```

```
##              Outcome +    Outcome -      Total                        Odds
## Exposed +            8           93        101        0.09 (0.03 to 0.16)
## Exposed -            4           94         98        0.04 (0.01 to 0.09)
## Total               12          187        199        0.06 (0.03 to 0.11)
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Exposure odds ratio                              2.02 (0.59, 6.94)
## Attrib fraction (est) in the exposed (%)      50.36 (-92.94, 89.43)
## Attrib fraction (est) in the population (%)   33.69 (-49.47, 70.58)
## -------------------------------------------------------------------
## Uncorrected chi2 test that OR = 1: chi2(1) = 1.294 Pr>chi2 = 0.255
## Fisher exact test that OR = 1: Pr>chi2 = 0.373
##  Wald confidence limits
##  CI: confidence interval
```

Prevalence ratio

	D+'ve	**D-'ve**	
Exp +'ve	a	b	
Exp -'ve	c	d	
			N

A cross sectional study investigating the relationship between dry cat food (DCF) and feline urologic syndrome (FUS) was conducted (Willeberg 1977). Counts of individuals in each group were as follows:

DCF-exposed cats (cases, non-cases) 13, 2163

Non DCF-exposed cats (cases, non-cases) 5, 3349

	FUS+'ve	FUS-'ve	
DCF +'ve	13	2,163	
DCF -'ve	5	3,349	
			5,530

```
epi.2by2(c(13,2163,5,3349), method="cross.sectional")
```

```
##              Outcome +    Outcome -      Total                Prev risk *
## Exposed +           13         2163       2176       0.60 (0.32 to 1.02)
## Exposed -            5         3349       3354       0.15 (0.05 to 0.35)
```

```
## Total               18         5512       5530       0.33 (0.19 to 0.51)
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Prev risk ratio                                4.01 (1.43, 11.23)
## Prev odds ratio                                4.03 (1.43, 11.31)
## Attrib prev in the exposed *                   0.45 (0.10, 0.80)
## Attrib fraction in the exposed (%)            75.05 (30.11, 91.09)
## Attrib prev in the population *                0.18 (-0.02, 0.38)
## Attrib fraction in the population (%)         54.20 (3.61, 78.24)
## -------------------------------------------------------------------
## Uncorrected chi2 test that OR = 1: chi2(1) = 8.177 Pr>chi2 = 0.004
## Fisher exact test that OR = 1: Pr>chi2 = 0.006
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 population units
```

We get a prevalence ratio of 4.01 (1.43-11.23) so we can say that the prevalence is 4 times higher in the exposed group compared to the unexposed group. Notice that the odds ratio is very similar. Also we can get an attributable prevalence (similar to the Attributable risk) which tells us that there is an additional probability of disease in the exposed group above baseline due to the exposure.

3.3 Measures of effect

For all these calculations we are working from a 2by2 table that is set out as below:

	D+'ve	**Dis-'ve**	
Exp +'ve	a	b	(n1 or a+b)
Exp -'ve	c	d	(n2 or c+d)

The RR, OR and PR are all extremely useful measures of association to help identify important factors and variables. But as we have said they do not reflect causation. Further, they do not capture the impact of the factor or the effect of carrying out an intervention (under the assumption that there is some causal link). Say for example that an exposure has a RR of 4, it means that if you have the exposure you are 4 times more likely to get the disease than if you do not have the exposure. However, if the exposure is very rare itself then you may not actually have much impact by trying to remove that exposure. There may be a factor that although it has a lower association because it is common could actually have much more impact in terms of reduction in disease. We shall try and demonstrate this with the example in a moment. First lets look at what these measures are. They essentually fall into 2 groups: (1) measures of effect in the expsoed group and (2) measures of effect at the population level.

Measures of effect in the exposed population In epidemiology, *attributable risk* is the difference in rate of a condition between an exposed population and an unexposed population. Attributable risk is mostly calculated in cohort studies, where individuals are studied on their exposure status and followed over a period of time. It also seems to be referred to as the risk difference or the incidence rate differnce in various textbooks.

$$attributable\, risk = \frac{a}{a+b} - \frac{c}{c+d}$$

or

$$attributable\ risk = \frac{a}{t1} - \frac{c}{t2}$$

So if we look at an example below

risk in the exposed group =

$$\frac{84}{84 + 2916} = 0.028$$

risk in the unexposed group =

$$\frac{87}{87 + 4923} = 0.017$$

$$attributable\ risk = 0.028 - 0.017 = 0.011$$

This indicates the increase in the probability of disease in the exposed group, beyond the baseline risk in the unexposed group.

The attributable fraction in comparison is the proportion of disease in the exposed individuals that is due to exposure.

$$attributable\ fraction = \frac{\frac{a}{a+b} - \frac{c}{c+d}}{\frac{a}{a+b}}$$

Thus in the example above

$$attributable\ risk = \frac{0.028 - 0.017}{0.028} = 0.393$$

In other words (assuming that the risk factor is causal) removing it would reduce the disease in the exposed group by 40%.

In summary

Attributable risk (AR) and attributable fraction (AF) both assess the impact of an exposure on an outcome, but they differ in focus and expression.

- **Attributable Risk (AR)**: Measures the absolute difference in risk or incidence between exposed and unexposed groups. It quantifies the actual number of cases per unit population attributable to the exposure, emphasizing the excess risk.
- **Attributable Fraction (AF)**: Expresses the proportion of the total incidence in the exposed group that can be attributed to the exposure. It is a relative measure, often calculated as $(\text{Risk}_{\text{exposed}} - \text{Risk}_{\text{unexposed}})/\text{Risk}_{\text{exposed}}$, highlighting the percentage of cases due to the exposure.

While AR focuses on magnitude, AF emphasizes proportion.

Measures of effect in the population

AR and AF above are useful for quantifying the effect of exposure in the exposed group but this still does not reflect te impact in the population as a whole. We can calculate both a *population attributable risk* and a *population attributable fraction* as below.

$$population\ attributable\ risk = \frac{a + c}{a + b + c + d} - \frac{c}{c + d}$$

In other words this is the overall risk of disease in the population minus that due to disease in the unexposed group.

The population attributable fraction reflects the effect of disease in the population rather than just the exposed group.

$$attributable\ fraction = \frac{\frac{a+c}{a+b+c+d} - \frac{c}{c+d}}{\frac{a+c}{a+b+c+d}}$$

or in other words the propotion of the disease in the population that is attributable to the exposure.

Let's have a look at a worked example to help explain these.

The preventive advantages of eating fish have been reported in numerous studies. A recent cohort study reported that not eating fish increased the risk for stroke. The table below shows the results of this study:

	Cases of Stroke	**Non Stroke**	
Never eat fish	82	1,549	1,631
Eat fish most days	23	779	802
	105	2,238	2,433

Incidence in the exposed (Ie):

$$\frac{a}{a+b} = \frac{82}{1,631} = 0.0503$$

Incidence in the unexposed (Iu):

$$\frac{c}{c+d} = \frac{23}{802} = 0.0287$$

Incidence in both combined (Ip):

$$\frac{a+c}{a+b+c+d} = \frac{105}{2,433} = 0.0432$$

$$Relative\ risk = \frac{0.0503}{0.0287} = 1.75$$

Applying the formulas above to these data (and disregarding the fact that some members of the population may eat fish more than "never" and less than "almost daily") results in the following measures of attributable risk.

$$Attributable\ risk = Ie - Iu = 5.03 - 2.87 = 2.16\ per\ 100$$

$$Attributable\ fraction = \frac{AR}{Ie} = \frac{2.16}{5.03} = 0.43\ or\ 43\%$$

$$Population\ attributable\ risk = Ip - Iu = 4.32 - 2.87 = 1.45\ per\ 100$$

$$Population\ attributable\ fraction = \frac{PAR}{Ip} = \frac{1.45}{4.32} = 0.336\ or\ 33.6\%$$

Assuming that this and many other studies present enough evidence about the preventive advantages of eating fish to reduce stroke, we could interpret the above data as follows:

- Those who never eat fish have 1.75 times the risk (higher incidence) as those who eat fish almost daily (RR = 1.75).

- If those who do not eat fish change their eating habits and begin to eat fish almost daily, their incidence of strokes will decrease by 2.16 per 100 individuals (AR = 2.16 per 100), which would represent a 43% reduction of their stroke incidence (AR% = 43%).
- A reduction of 1.45 new cases of stroke per 100 population (exposed and unexposed) is expected if everybody eats fish almost daily (PAR = 1.45 per 100). Such reduction represents a 33.6% reduction of the incidence in the population (PAR% = 33.6%).

Reference Sauvaget C, Nagano J, Allen N, et al. Intake of animal products and stroke mortality in the Hiroshima/Nagasaki Life Span Study. International Journal of Epidemiology. 2003;32:536–543.

Exercises

Ex 3.1.

You set out and do a study of 90 cattle randomly selected from a herd of 320 and tested for antibodies to EBL. The farmer is concerned that buying-in animals might be a risk factor for EBL in the herd. A study of the records showed that 19 of the study animals had been bought in, whereas 71 had been bred on the farm with EBL status as follows:

	EBL+'ve	**EBL-'ve**
Bought in	12	7
Home bred	35	36

a) what sort of study is this?
b) estimate the association between the exposure (with 95% CI) and being EBL positive using the appropriate estimator showing your working (or code in R)
c) write a brief paragraph explaining the results

Ex 3.2.

A telephone questionnaire was used to gather information from owners of 38 cats diagnosed at the Small Animal Hospital with chronic renal failure (CRF) and from 56 cats visiting the clinic for annual vaccination boosters. Data were collected from all owners on how they fed their cats to determine whether diet was a risk factor for chronic renal failure. The following data were recorded:

	CRF+'ve	**CRF-'ve**
ad lib +'ve	35	41
ad lib -'ve	3	15

a) what sort of study is this?
b) estimate the association between the exposure (with 95% CI) and having CRF using the appropriate estimator showing your working (or code in R)
c) write a brief paragraph explaining the results "'