Measures of association and effect

Hynek Pikhart

Revision - Measures of disease frequency

- Used for binary outcomes
- Require a numerator and denominator

$$\frac{\text{number of persons with disease}}{\text{number of persons examined}}$$

- expressed as X per 1000 persons (or per 100,000 etc)

Prevalence

- number of **existing** cases / population of interest at a defined time

Incidence

- number of **new** cases in a given time period / total population at risk

Measures of association

- Risk of disease, rate of disease in different groups of population
- Comparison of risks/rates

Constructing 2-way table

For binary health outcomes (Y/N), it is possible to construct 2x2 table and to estimate either relative or absolute measures of risk

	Disease		
Exposure	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Relative measures of effect (relative risk)

We have 2 groups of individuals:

- An **exposed** group (group with risk factor of interest) and **unexposed** group (without such factor of interest)
- We are interested in <u>comparing</u> the amount of disease (mortality or other health outcome) in the exposed group to that in the unexposed group

Risk/rate

- Incidence rate or Risk in exposed (r_1)
- Incidence rate or Risk in unexposed (r_0)

Measures of association

- Risk of disease, rate of disease in different groups of population
- Comparison of risks/rates

Risk ratio

- we calculate the risk ratio (RR) as:

$$RR=r_1/r_0$$

Risk difference

- the absolute difference between two risks (or rates)

$$RD = r_1 – r_0$$

Constructing 2-way table

For binary health outcomes (Y/N), it is possible to construct 2x2 table and to estimate either relative or absolute measures of risk

	Disease		
Exposure	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Example: Alcohol drinking and heart attack

	Heart attack		
	Yes	No	Total
Alcohol drinking			
Yes	25	400	425
No	75	1500	1575
Total	100	1900	2000

Risk (exposed) = 25/425=0.059
Risk (unexposed) = 75/1575=0.048

Relative risk = 0.059/0.048 = 1.23

- We can also have different strata of exposure. We may calculate ratio measures for each strata – we compare measure of frequency in each level with measure of frequency in the baseline (unexposed) level.
- *Example: Death rates from CHD in smokers and non-smokers by age*

Age	Smokers rate	Non-smokers rate	**Rate ratio**
35-44	0.61	0.11	**5.5**
45-54	2.40	1.12	**2.1**
55-64	7.20	4.90	**1.5**
65-74	14.69	10.83	**1.4**
75-84	19.18	21.20	**0.9**
85+	35.93	32.66	**1.1**
ALL AGES	**4.29**	**3.30**	**1.3**

What can you say about this table?

Age	Smokers rate	Non-smokers rate	Rate ratio
35-44	0.61	0.11	**5.5**
45-54	2.40	1.12	**2.1**
55-64	7.20	4.90	**1.5**
65-74	14.69	10.83	**1.4**
75-84	19.18	21.20	**0.9**
85+	35.93	32.66	**1.1**
ALL AGES	**4.29**	**3.30**	**1.3**

The rate ratio decreases with increasing age. This table may also suggest that the effect of smoking on the rate of CHD is higher in younger ages.

Odds ratio

- Alternative measure of risk

The odds of disease is the number of cases divided by the number of non-cases

$$\text{Odds} = \frac{\text{Cases}}{\text{Non cases}}$$

Odds ratio (**OR**) is ratio of odds of disease among exposed ($odds_{exp}$) and odds of disease among unexposed ($odds_{unexp}$)

$$\mathbf{OR = odds_{exp} / odds_{unexp}}$$

	Heart attack		
	Yes	No	Total
Alcohol drinking			
Yes	25	400	425
No	75	1500	1575
Total	100	1900	2000

We can calculate

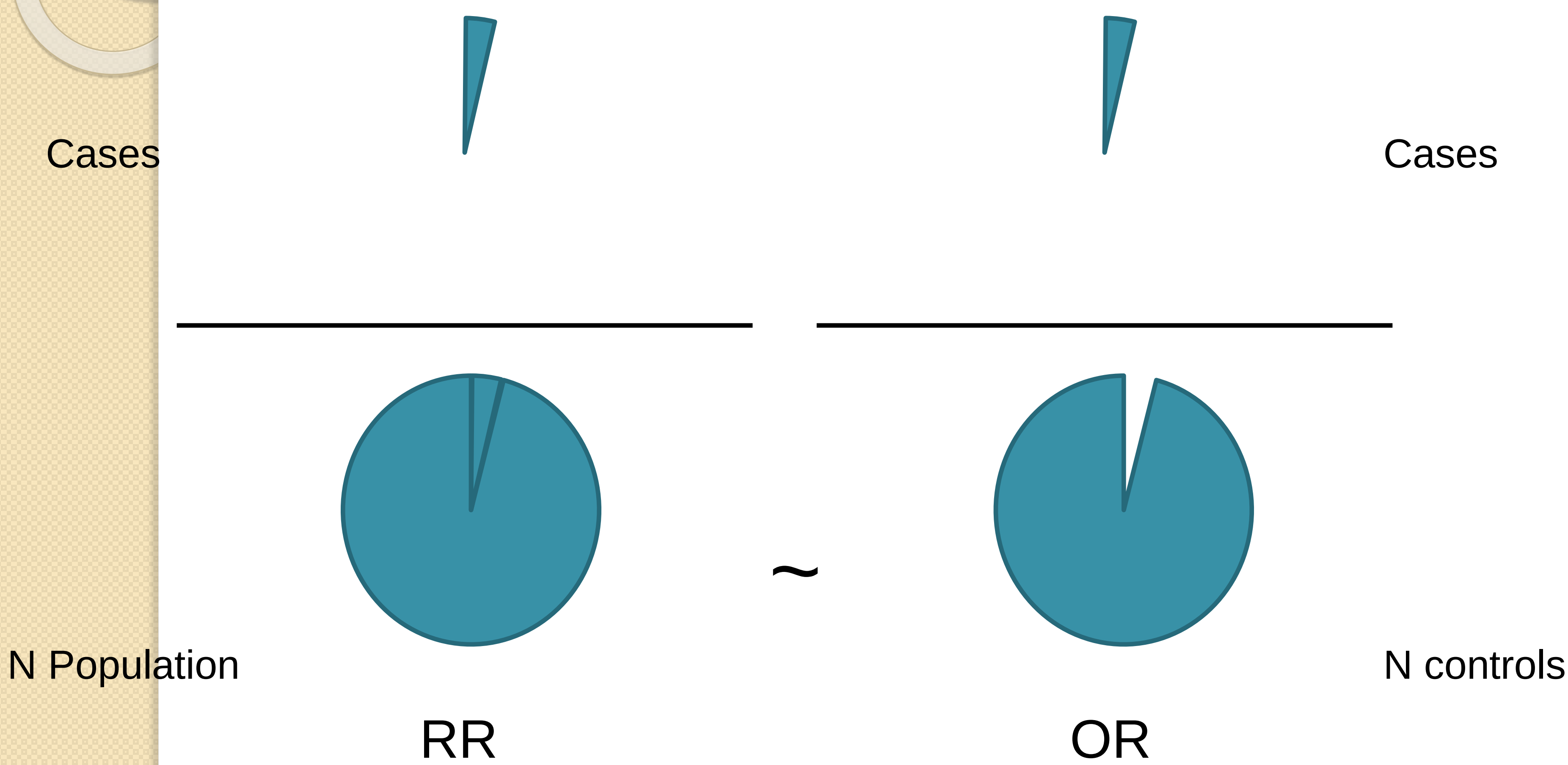
- Odds (exposed) O_{exp}=25/400
- Odds (unexposed) O_{unexp}=75/1500

- Odds ratio $OR = O_{exp} / O_{unexp}$ = 1.25

Odds ratio as an approximation to the risk ratio

- For a rare disease, odds ratio is approximately equal to the risk ratio (because denominators are very similar)
- For a common conditions, OR overestimates the true RR

Rare disease → OR~RR



If disease common:

Disease	Exposed	Unexposed	Total
Yes	50	25	75
No	50	75	125
Total	100	100	200

$\frac{a / (a+b)}{c / (c+d)}$ R_1=50/100=0.5 R_0=25/100=0.25 RR=2.0

$\frac{a / b}{c / d}$ O_1=50/50=1.0 O_0=25/75=0.33 OR=3.0

<table>
<tr><th>Measure of effect</th><th>Use of the measure</th><th>How to interpret results</th></tr>
<tr><td>Risk Difference</td><td>Public Health
Interested in excess disease burden due to factor (“Attributable risk”)</td><td>Close to 0 = little effect
Large difference = large effect</td></tr>
<tr><td>Risk Ratio</td><td>Epidemiology
Causation
“This factor doubles the risk of the disease”</td><td rowspan="2">Close to 1 = little effect
Large ratio = large effect
Close to 0 = large effect!</td></tr>
<tr><td>Odds Ratio</td><td>As for Risk Ratio
“This factor doubles the odds of the disease”
Only possibility (case-control study)
More advanced statistical methods (logistic regression)</td></tr>
</table>

Example

- Random sample of individuals were questioned about their occupation and their BP was measured. Based on SBP and DBP measures they were classified as hypertensive or non-hypertensive. Among 300 people in non-manual jobs, there were 72 hypertensive individuals. Among 240 people in manual jobs, there were 96 hypertensive individuals.

Constructing 2-way table

As a first step we need to organize our data in a formal way – we construct 2-way table

	Hypertension		
	Yes	No	Total
Manual	96	144	240
Non-manual	72	228	300
Total	168	372	540

What does it mean when we speak about an association between two categorical variables?

- It means that knowing the value of one variable tells us something about the value of the other variable.
- Two variables are therefore said to be associated if the distribution of one variable varies according to the value of the other variable.

What does it mean when we speak about an association between two categorical variables?

- In our example, the two variables, occupation and hypertension, are associated if the distribution of hypertension varies between occupational groups.
- And, if distribution of hypertension is same in both occupational groups, we can say that there is no association between hypertension and occupational category - because knowing a occupational category of individual will not tell us anything about hypertension.

What does it mean when we speak about an association between two categorical variables?

- Having constructed a two-way table, the next step is to look whether the distribution of one variable differs according to the value of the other variable.
- We need to calculate either row or column percentages.
- Often, one variable can be regarded as **the response variable**, while the other is **the explanatory variable**, and this should help to decide what percentages are shown
- If the columns represent the explanatory variable, then column percentages are more appropriate, and vice versa.

Constructing 2-way table

As a second step we calculate proportion of hypertensive individuals among manual workers, non-manual workers and in the whole sample

	Hypertension		
	Yes	No	Total
Manual	96 (40.0%)	144 (60.0%)	240
Non-manual	72 (24.0%)	228 (76.0%)	300
Total	168 (31.1%)	372 (68.9%)	540

The numbers in the four categories in the 2-way table in the previous slide all called

OBSERVED NUMBERS

- The data seem to suggest some association between hypertension and occupation (40% of manual workers with hypertension compared to 24% of non-manual workers with hypertension)

- The calculation and examination of such percentages is an essential step in the analysis of a two-way table, and should always be done before starting formal significance tests.

Significance test for the association

- Although it seems that there is an association in the table, the question is whether this may be attributable to sampling variability
- Each of the percentages in the table is subject to sampling error, and we need to assess whether the differences between them may be due to chance
- This is done by conducting a significance test
- The null hypothesis is "***there is no association between the two variables***"

Expected numbers

- The significance test is **Chi-squared test**

- This test compares the **observed** numbers in each of four categories of contingency table with the numbers to be **expected** if there was no difference in proportion of hypertensive individuals in two occupational groups

Expected numbers

	Hypertension		
	Yes	No	Total
Manual	**74.64**		240
Non-manual			300
Total	168 (31.1%)	372 (68.9%)	540

- From the table above, the overall proportion of hypertensive individuals is 168/372 (31.1%).
- If the null hypothesis were true, the expected number of manual subjects with hypertension is 31.1% of 240, which is 74.64

- Expected numbers in the other cells of the table can be calculated similarly, using the general formula:

$$\textbf{Expected number} = \frac{\textbf{Row total} \times \textbf{Column total}}{\textbf{Overall total}}$$

	Hypertension		
	Yes	No	Total
Manual	74.64	165.36	240
Non-manual	93.36	206.64	300
Total	168 (31.1%)	372 (68.9%)	540

Next step – compare observed and expected numbers

OBSERVED	Hypertension		
	Yes	No	Total
Manual	96	144	240
Non-manual	72	228	300
Total	168 (31.1%)	372 (68.9%)	540

EXPECTED	Hypertension		
	Yes	No	Total
Manual	74.64	165.36	240
Non-manual	93.36	206.64	300
Total	168 (31.1%)	372 (68.9%)	540

Chi-squared test (X^2 test)

$$X^2 = \Sigma \left[(O - E)^2/E\right]$$

- Calculate $(O-E)^2/E$ for each cell and sum over all cells

- In our example:

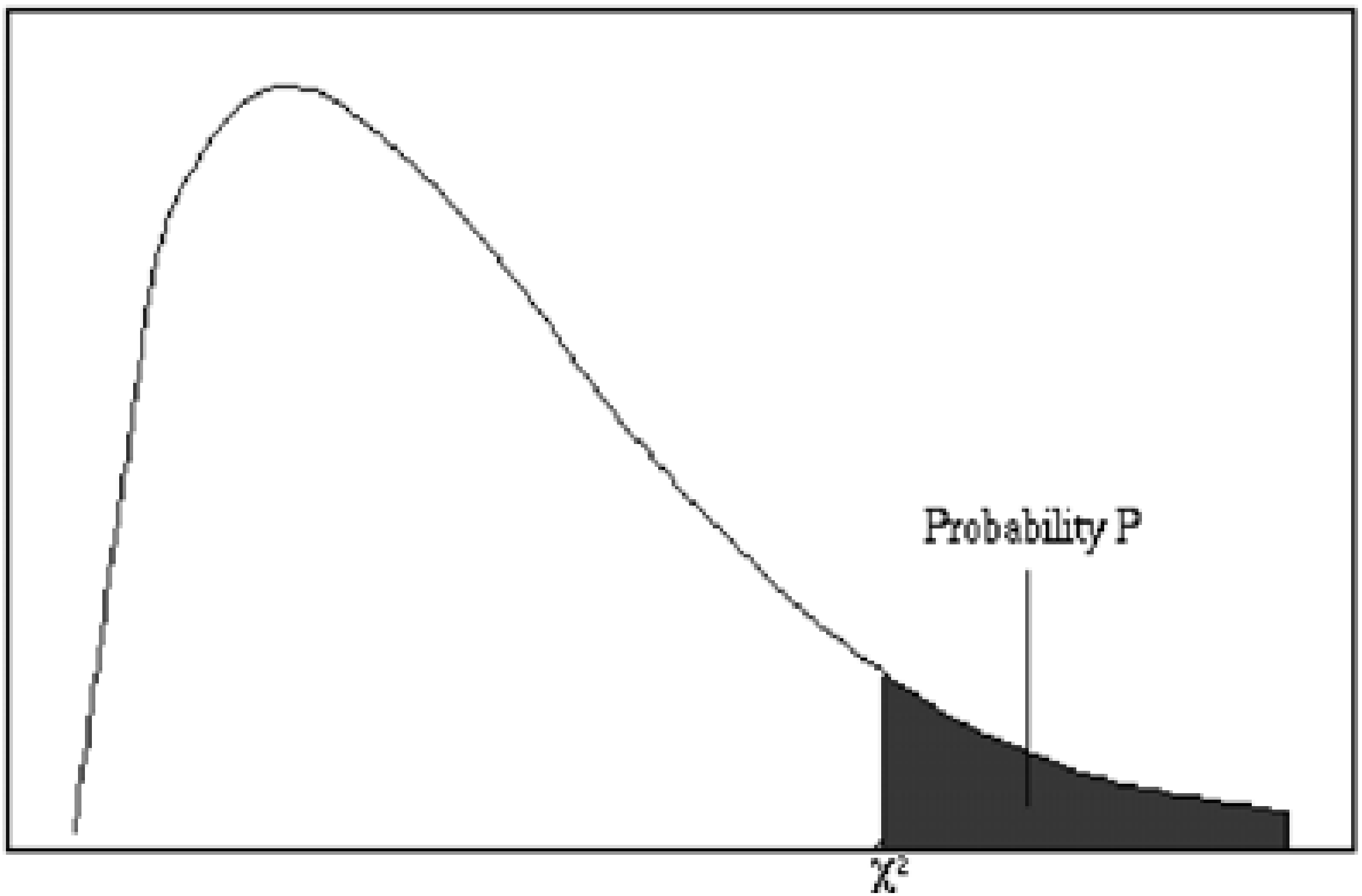
$X^2 = [(96-74.64)^2 / 74.64 + (144-165.36)^2 / 165.36 + (72-93.36)^2 / 93.36 + (228-206.64)^2 / 206.64] =$ **15.97**

- If χ^2 value is large then (O-E) is, in general, large and data do not support H_0 = **association**
- If χ^2 value is small then (O-E) is, in general, small and data do support H_0 = **no association**

- Large values of χ^2 suggest that the data are **inconsistent** with the null hypothesis, and therefore that there is an association between the two variables.

Obtaining p-value

- Under H_0: $\chi 2$ distribution



Obtaining p-value

- The P-value is obtained by referring the calculated value of **χ**2 to tables of the chi-squared distribution.
- The P-value in this case corresponds to the value shown as α in the tables.
- The degrees of freedom are given by the formula:

$$d.f. = (r - 1) \times (c - 1)$$

- r = number of rows, c = number of columns

Table I. Critical Values of χ^2

	Level of Significance for Two-tailed Test					
df	.20	.10	.05	.02	.01	.001
1	1.64	2.71	3.84	5.41	6.64	10.83
2	3.22	4.60	5.99	7.82	9.21	13.82
3	4.64	6.25	7.82	9.84	11.34	16.27
4	5.99	7.78	9.49	11.67	13.28	18.46
5	7.29	9.24	11.07	13.39	15.09	20.52
6	8.56	10.64	12.59	15.03	16.81	22.46
7	9.80	12.02	14.07	16.62	18.48	24.32
8	11.03	13.36	15.51	18.17	20.09	26.12
9	12.24	14.68	16.92	19.68	21.67	27.88
10	13.44	15.99	18.31	21.16	23.21	29.59
11	14.63	17.28	19.68	22.62	24.72	31.26
12	15.81	18.55	21.03	24.05	26.22	32.91
13	16.98	19.81	22.36	25.47	27.69	34.53
14	18.15	21.06	23.68	26.87	29.14	36.12
15	19.31	22.31	25.00	28.26	30.58	37.70
16	20.46	23.54	26.30	29.63	32.00	39.29
17	21.62	24.77	27.59	31.00	33.41	40.75
18	22.76	25.99	28.87	32.35	34.80	42.31
19	23.90	27.20	30.14	33.69	36.19	43.82
20	25.04	28.41	31.41	35.02	37.57	45.32
21	26.17	29.62	32.67	36.34	38.93	46.80
22	27.30	30.81	33.92	37.66	40.29	48.27
23	28.43	32.01	35.17	38.97	41.64	49.73
24	29.55	33.20	36.42	40.27	42.98	51.18
25	30.68	34.38	37.65	41.57	44.31	52.62
26	31.80	35.56	38.88	42.86	45.64	54.05
27	32.91	36.74	40.11	44.14	46.96	55.48
28	34.03	37.92	41.34	45.42	48.28	56.89
29	35.14	39.09	42.69	46.69	49.59	58.30
30	36.25	40.26	43.77	47.96	50.89	59.70
32	38.47	42.59	46.19	50.49	53.49	62.49
34	40.68	44.90	48.60	53.00	56.06	65.25
36	42.88	47.21	51.00	55.49	58.62	67.99
38	45.08	49.51	53.38	57.97	61.16	70.70
40	47.27	51.81	55.76	60.44	63.69	73.40
44	51.64	56.37	60.48	65.34	68.71	78.75
48	55.99	60.91	65.17	70.20	73.68	84.04
52	60.33	65.42	69.83	75.02	78.62	89.27
56	64.66	69.92	74.47	79.82	83.51	94.46
60	68.97	74.40	79.08	84.58	88.38	99.61

Back to our example:

$X^2 = 15.97$

Table 2x2

d.f.=1

and from the table

$P<0.001$

Larger tables (r x c tables)

$$X^2 = \sum \frac{(O-E)^2}{E}$$

d.f. = (r-1) x (c-1)

- Valid if less than 20% of expected numbers are under 5 and none is less than 1
- If low expected numbers – combine either rows or columns to overcome this problem

How to calculate expected number in particular cell

$$\text{Expected number} = \frac{\text{Row total} \times \text{Column total}}{\text{Overall total}}$$

Interpretation of chi-square test results: Chi-squared tests in STATA

- We try to evaluate whether there is an association between current smoking and age
- We have age grouped into 4 groups (30-39, 40-49, 50-59, 60-69)
- Smoking (variable smok) was coded 1=current smokers, 0=non-smokers

Let's check proportion of smokers in each age category

```
. tab smok agegroup, col

             |          30-39,40-49,50-59,60-69
1=yes 0=no   |        30         40         50         60 |     Total
-------------+--------------------------------------------+----------
           0 |       337        357        490        491 |     1,675
             |     54.71      56.31      72.38      78.81 |     65.69
-------------+--------------------------------------------+----------
           1 |       279        277        187        132 |       875
             |     45.29      43.69      27.62      21.19 |     34.31
-------------+--------------------------------------------+----------
       Total |       616        634        677        623 |     2,550
             |    100.00     100.00     100.00     100.00 |    100.00
```

Chi-squared test

```
. tab smok agegroup, col chi

            |       30-39,40-49,50-59,60-69
1=yes 0=no |        30         40         50         60 |     Total
-----------+--------------------------------------------+----------
         0 |       337        357        490        491 |     1,675
           |     54.71      56.31      72.38      78.81 |     65.69
-----------+--------------------------------------------+----------
         1 |       279        277        187        132 |       875
           |     45.29      43.69      27.62      21.19 |     34.31
-----------+--------------------------------------------+----------
     Total |       616        634        677        623 |     2,550
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) =  118.7458   Pr = 0.000
```

Degrees of freedom **Chi-squared test value** **p<0.001**

Measures of population impact

- **Population attributable risk (PAR)** is the absolute difference between the risk (or rate) in the whole population and the risk or rate in the unexposed group

$$PAR = r - r_0$$

Population attributable risk fraction (PARF or PAR%)

- It is a measure of the proportion of all cases in the study population (exposed and unexposed) that may be attributed to the exposure, on the assumption of a causal association
- It is also called the aetiologic fraction, the percentage population attributable risk or the attributable fraction

- If r is rate in the total population

$$\mathbf{PAF = PAR/r}$$

$$\mathbf{PAR = r - r_0}$$

$$\mathbf{PAF = (r\text{-}r_0)/r}$$

Exercise

- 50 persons attended a garden party
- 25 of them developed diarrhoea in the next 3 days
- What was the risk of diarrhoea among the participants of the party?

Exercise – cont.

- 30 party visitors had a BBQ (minced meat)
- 24 of them developed diarrhoea

- 20 people did not eat BBQ
- 1 of them developed diarrhoea

- How would you calculate RR related to eating BBQ?

Exercise – cont.

- Risk among unexposed R_0:
- 1/20

- Risk among exposed R_1:
- 24/30

- Relative risk $RR=R_1/R_0=(24/30)/(1/20)=16$