Chapter 9: Contingency analysis – associations between categorical variables

1

Contingency analysis

- **Contingency analysis** estimates and tests for an association between two or more categorical variables
- Reminder: Categorical data are qualitative characteristics of individuals that do not have magnitude on a numerical scale
  - e.g., present or absent; language (English, Spanish, etc.); life stage (egg, larvae, adult)
- Determines the extent to which one variable is "contingent" on the other

2

Contingency analysis examples

- Do bright and drab butterflies have the same probability of being eaten?
  - color (bright or drab); eaten (yes or no)
- Are smokers more likely to drink than non-smokers?
  - drink (yes or no); smoke (yes or no)
- Were men and women equally likely to survive the sinking of the Titanic?
  - sex (male or female); survive (yes or no)

3

2 x 2 contingency tables

- Two categorical variables, each with two groups
- Were men and women equally likely to survive the sinking of the Titanic?
  - sex (male or female); survive (yes or no)

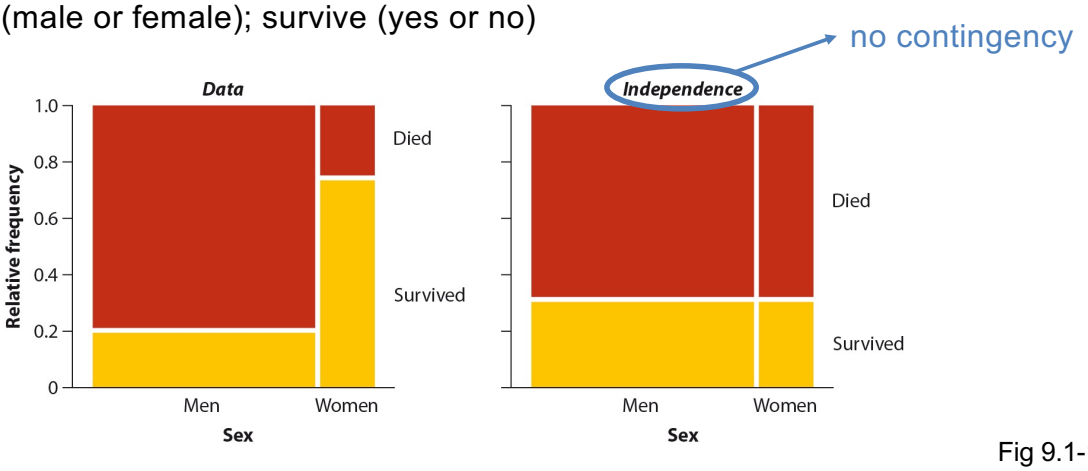


Fig 9.1-1

4

2 x 2 contingency tables

- Three different ways to analyze these data
  - Relative risk
  - Odds ratio
  - χ^2 contingency test

5

2 x 2 contingency tables

- Three different ways to analyze these data
  - Relative risk
  - Odds ratio
  - χ^2 contingency test

6

2 x 2 contingency: relative risk

- **Relative risk** is the probability of an undesired outcome in the treatment group divided by the probability of the same outcome in a control group

Parameter

$$RR = \frac{p_1}{p_2}$$

$RR: relative\ risk$
$p: probability\ of\ undesired\ outcome$

Estimate/Statistic

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2}$$

$\widehat{RR}: estimate\ of\ relative\ risk$
$\hat{p}: observed\ probability\ of\ undesired\ outcome\ in\ treatment\ (1) or\ control\ (2) groups$

7

Ex 9.2: Take two aspirin and call me in the morning?

- Aspirin has been shown to reduce the risk of stroke and heart attack
- Does it also reduce the risk of cancer?
- 10 year study in which 39,876 woman took either aspirin or placebo every other day

TABLE 9.2-1 2 × 2 contingency table for the aspirin and cancer experiment.

	Aspirin	Placebo
Cancer	1438	1427
No cancer	18,496	18,515

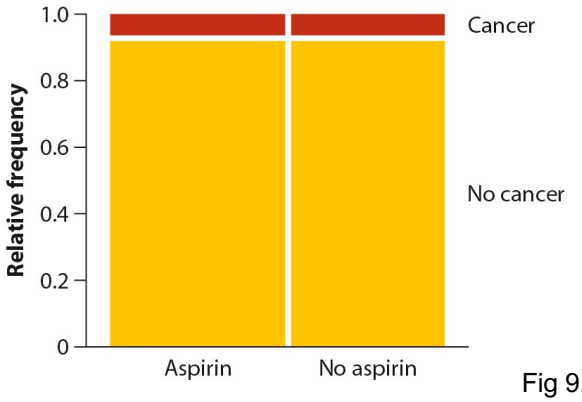


Fig 9.2-1

8

Relative risk calculations

	Aspirin	Placebo
Cancer	$a = 1438$	$b = 1427$
No cancer	$c = 18{,}496$	$d = 18{,}515$
	Sum: 19,934	19,942

In medical studies, the convention is:

- *to calculate the odds of the outcome "diseased" or "died"*
- *Set the control/placebo as group 2*

$$\hat{p}_1 = \frac{1438}{19934} = 0.0721$$

$$\hat{p}_2 = \frac{1427}{19942} = 0.0716$$

$$\widehat{RR} = \frac{0.0721}{0.0716} = 1.007$$

Note: results not exact due to rounding

9

Relative risk SE and confidence interval

Standard error of relative risk → measure of uncertainty

$$SE\left[ln\left(\widehat{RR}\right)\right] = \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{a+c} - \frac{1}{b+d}}$$

95% confidence interval of relative risk → 95% conf true parameter lies in this range

$$ln\left(\widehat{RR}\right) - 1.96\left(SE\left[ln\left(\widehat{RR}\right)\right]\right) < ln(RR) < ln\left(\widehat{RR}\right) + 1.96\left(SE\left[ln\left(\widehat{RR}\right)\right]\right)$$

Note: sampling distribution of the relative risk ratio is highly skewed, so a conversion to the natural log is used

10

Relative risk SE and confidence interval

Standard error of relative risk

$$SE\left[ln\left(\widehat{RR}\right)\right] = \sqrt{\frac{1}{a}+\frac{1}{b}-\frac{1}{a+c}-\frac{1}{b+d}} = \sqrt{\frac{1}{1438}+\frac{1}{1427}-\frac{1}{1438+18496}-\frac{1}{1427+18515}} = 0.03600$$

95% confidence interval of relative risk

$$ln\left(\widehat{RR}\right) - 1.96\left(SE\left[ln\left(\widehat{RR}\right)\right]\right) < ln(RR) < ln\left(\widehat{RR}\right) + 1.96\left(SE\left[ln\left(\widehat{RR}\right)\right]\right)$$

$$0.00698 - 1.96(0.03600) < ln(RR) < 0.00698 + 1.96(0.03600)$$

$$-0.064 < ln(RR) < 0.078$$

$$e^{-0.064} < RR < e^{0.078}$$

$$0.94 < RR < 1.08$$

11

Conclusion based on relative risk

- $0.93 < RR < 1.09$
- Potential outcomes
  - 1: risk in two groups are equal
  - <1: risk in group 2 (denominator, placebo/control) more likely
  - >1: risk in group 1 (numerator, treatment) more likely
- Range includes 1 (no difference in risk of getting cancer between aspirin and placebo groups)
- Data are consistent with a small beneficial effect, a small deleterious effect, or no effect at all

12

2 x 2 contingency tables

- Three different ways to analyze these data
  - Relative risk
  - Odds ratio
  - χ^2 contingency test

13

2 x 2 contingency: odds ratio

- The odds of success are the probability of success divided by the probability of failure

Parameter

$$O = \frac{p}{1-p}$$

O: odds of success
p: probabality of success

Estimate/Statistic

$$\hat{O} = \frac{\hat{p}}{1-\hat{p}}$$

$\hat{O}$: estimate of odds of success
$\hat{p}$: observed proportion of successes

If O = 1 then there the odds of successes and failures are equal, thus no contingency between variables

14

Ex 9.2: Take two aspirin and call me in the morning?

- Aspirin has been shown to reduce the risk of stroke and heart attack
- Does it also reduce the risk of cancer?
- 10 year study in which 39,876 woman took either aspirin or placebo every other day

TABLE 9.2-1 2 × 2 contingency table for the aspirin and cancer experiment.

	Aspirin	Placebo
Cancer	1438	1427
No cancer	18,496	18,515

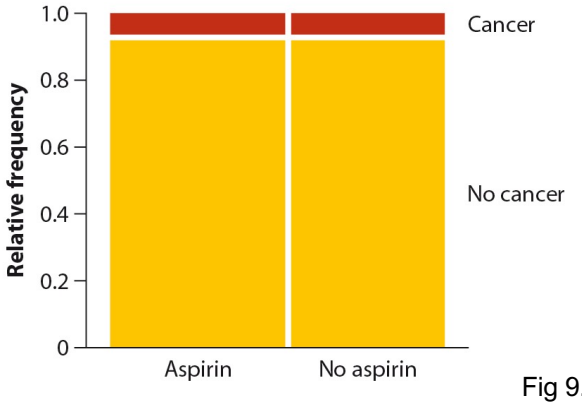


Fig 9.2-1

15

Ex 9.2: Take two aspirin and call me in the morning?

TABLE 9.2-1 2 × 2 contingency table for the aspirin and cancer experiment.

	Aspirin	Placebo
Cancer	1438	1427
No cancer	18,496	18,515
	Sum: 19,934	19,942

Group 1: Aspirin

$$\hat{p}_1 = \frac{1438}{19934} = 0.0721$$

$$\hat{O}_1 = \frac{\hat{p}_1}{1-\hat{p}_1} = \frac{0.0721}{0.9279} = 0.0777$$

Group 2: No aspirin

$$\hat{p}_2 = \frac{1427}{19942} = 0.0716$$

$$\hat{O}_2 = \frac{\hat{p}_2}{1-\hat{p}_2} = \frac{0.0716}{0.9284} = 0.0771$$

16

Testing for difference between groups

- Odds of developing cancer for aspirin group
  - $\hat{O}_1 = 0.0777$
- Odds of developing cancer in no aspirin (placebo) group
  - $\hat{O}_2 = 0.0771$
- Are these statistically different?

17

Odds ratio

- The odds ratio is the odds of success in one group divided by the odds of success in a second group

$$\widehat{OR} = \frac{\hat{O}_1}{\hat{O}_2}$$

The convention in medical studies is to make the control/placebo group #2 (denominator)

- Potential outcomes
  - 1: odds of success in two groups are equal
  - <1: odds of success (disease) in group 1 (numerator, treatment) less likely
  - >1: odds of success in group 1 (numerator, treatment) more likely

18

Odds ratio calculation shortcut

	Treatment	Control
Success (focal outcome)	*a*	*b*
Failure (alternative outcome)	*c*	*d*

$$\widehat{OR} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

	Aspirin	Placebo
Cancer	*a* = 1438	*b* = 1427
No cancer	*c* = 18,496	*d* = 18,515

$$\widehat{OR} = \frac{ad}{bc} = \frac{(1438)(18515)}{(1427)(18496)} = 1.009$$

19

Odds ratio SE and confidence interval

Standard error of odds ratio → measure of uncertainty

$$SE\left[ln\left(\widehat{OR}\right)\right] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

95% confidence interval → 95% conf true parameter lies in this range

$$ln\left(\widehat{OR}\right) - 1.96\left(SE\left[ln\left(\widehat{OR}\right)\right]\right) < ln(OR) < ln\left(\widehat{OR}\right) + 1.96\left(SE\left[ln\left(\widehat{OR}\right)\right]\right)$$

Note: sampling distribution of the odds ratio is highly skewed, so a conversion to the natural log is used

20

Odds ratio SE and confidence interval

Standard error of odds ratio

$$SE\left[ln\left(\widehat{OR}\right)\right] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{1438}+\frac{1}{1427}+\frac{1}{18496}+\frac{1}{18515}} = 0.03878$$

95% confidence interval

$$ln\left(\widehat{OR}\right) - 1.96\left(SE\left[ln\left(\widehat{OR}\right)\right]\right) < ln(OR) < ln\left(\widehat{OR}\right) + 1.96\left(SE\left[ln\left(\widehat{OR}\right)\right]\right)$$

$$0.00896 - 1.96(0.03878) < ln(OR) < 0.00896 + 1.96(0.03878)$$

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

$$e^{-0.067} < OR < e^{0.085}$$

$$0.93 < OR < 1.09$$

21

Conclusion

- 95% confident that OR is between 0.93 and 1.09
- Range includes 1 (no difference in odds of getting cancer between aspirin and placebo groups)
- Data are consistent with a small beneficial effect, a small deleterious effect, or no effect at all

22

Odds ratio or relative risk?

- Both can be used when there are two categorical variables, each with two groups (i.e., 2 x 2 table)
- Both are often used in biomedical studies

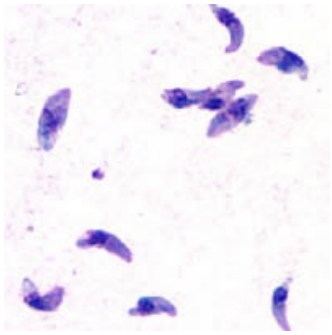
23

Odds ratio or relative risk?

- $\widehat{RR}$ plus: more intuitive (ratio of two proportions)
- $\widehat{OR}$ plus: can be applied to data from case-control studies
- A **case-control study** is a type of observational study in which a sample of individuals with a focal condition (cases) is compared to a sample of subjects lacking the condition (controls)
  - Sample sizes set by experimenter, thus proportion of individuals with/without condition is not proportional to the population parameters

24

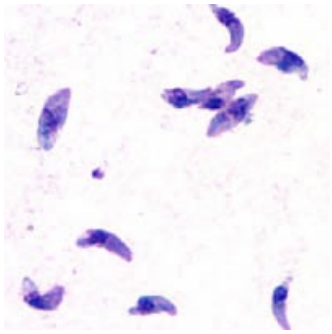
Ex 9.3: Your litter box and your brain



- *Toxoplasma gondii* – protozoan parasite
  - Primary host = cats
- Shown to affect brains/behavior in rats and mice
  - Reduced fear of cats
- ~25% of humans are infected
- May be associated with mental illness and increased risky behavior in humans
- Hypothesis: toxoplasmosis changes the probability of getting into a car accident

25

Ex 9.3: Your litter box and your brain



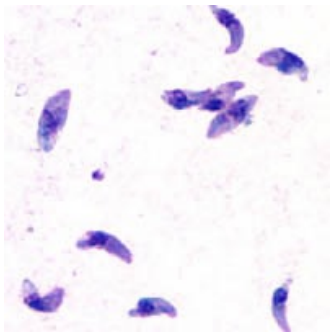
- Hypothesis: toxoplasmosis changes the probability of getting into a car accident
- Case-control study
- Explanatory variable = toxoplasmosis

Cases: Car accident	**Controls: No car accident**
185 individuals	185 individuals
21-40 years old	21-40 years old

Test for different probability of toxoplasmosis in two groups

26

Ex 9.3: Your litter box and your brain



- Cannot estimate the relative risk
- Do not have a random sample of drivers to estimate the probability of getting into an accident
- Data are enriched with drivers that have had an accident compared to a random sample from the population
- **But can estimate the odds ratio**

27

Ex 9.3: Your litter box and your brain

TABLE 9.3-1 The frequency of *Toxoplasma gondii* infection in a sample of drivers involved in driving accidents (cases) compared with a sample of drivers with no accidents (controls). From Yereli et al. (2006).

	Infected	Uninfected
Drivers with accidents	61	124
Drivers without accidents	16	169

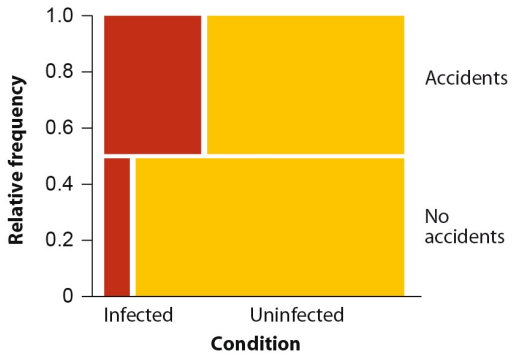


Fig 9.3-1

28

Odds ratio

	Treatment	Control
Success (focal outcome)	*a*	*b*
Failure (alternative outcome)	*c*	*d*

$$\widehat{OR} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

TABLE 9.3-1 The frequency of *Toxoplasma gondii* infection in a sample of drivers involved in driving accidents (cases) compared with a sample of drivers with no accidents (controls). From Yereli et al. (2006).

	Infected	Uninfected
Drivers with accidents	61	124
Drivers without accidents	16	169

$$\widehat{OR} = \frac{ad}{bc} = \frac{(61)(169)}{(16)(124)} = 5.20$$

95% confidence interval:

$$2.86 < OR < 9.44$$

29

Odds ratio

TABLE 9.3-1 The frequency of *Toxoplasma gondii* infection in a sample of drivers involved in driving accidents (cases) compared with a sample of drivers with no accidents (controls). From Yereli et al. (2006).

	Infected	Uninfected
Drivers with accidents	61	124
Drivers without accidents	16	169

95% confidence interval:

$$2.86 < OR < 9.44$$

- 95% confidence interval **does not** include 1
- Odds of getting into an accident are considerably (2.9-9.4x) higher for people with toxoplasmosis!

30

2 x 2 contingency tables

- Three different ways to analyze these data
  - Odds ratio
  - Relative risk
  - χ^2 contingency test

31

χ^2 contingency test

- The **χ^2 contingency test** is the most commonly used test of association between two categorical variables
- Special case of the χ^2 goodness-of-fit test where the null model is independence of variables
- Same assumptions of χ^2 goodness-of-fit test
  - No cells have expected frequency less than one
  - No more than 20% of cells have expected frequency less than five
- If these assumptions are not met then can use Fisher's exact test

32

Ex 9.4: The gnarly worm gets the bird

- Parasite *Euhaplorchis californiensis* uses three different hosts during its life cycle
  - Bird
  - Snail
  - Fish
- Observation that infected fish spend more time near surface
- Are infected fish more likely to be eaten by birds?

33

Ex 9.4: The gnarly worm gets the bird

- Outdoor tank with uninfected, lightly infected, and highly infected fish
- Recorded number of fish eaten

TABLE 9.4-1 Observed frequencies of fish eaten or not eaten by birds according to trematode infection level.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	1	10	37	48
Not eaten by birds	49	35	9	93
Column total	50	45	46	141

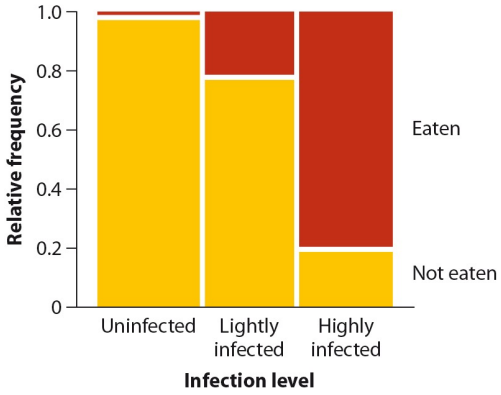


Fig 9.4-1

34

Hypotheses

- H_0: Parasite infection and being eaten are independent
- H_A: Parasite infection and being eaten are not independent

35

Expected frequencies?

TABLE 9.4-1 Observed frequencies of fish eaten or not eaten by birds according to trematode infection level.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	1	10	37	48
Not eaten by birds	49	35	9	93
Column total	50	45	46	141

TABLE 9.4-2 Expected frequencies of fish eaten and not eaten by birds, according to trematode infection status.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	17.0			48
Not eaten by birds				93
Column total	50	45	46	141

Pr[*uninfected* and *eaten*]

Assume independence-multiplication rule!

$$\widehat{Pr}[uninfected] = \frac{total\ uninfected}{total\ sample}$$

$$\widehat{Pr}[uninfected] = \frac{50}{141} = 0.3546$$

$$\widehat{Pr}[eaten] = \frac{total\ eaten}{total\ sample}$$

$$\widehat{Pr}[uninfected] = \frac{48}{141} = 0.3404$$

$$\widehat{Pr}[uninf\ and\ eaten] = 0.3546 \times 0.3404 = 0.1207$$

$$Expected[uninf\ and\ eaten] = 0.1207 \times 141 = 17.0$$

36

Expected frequencies

TABLE 9.4-1 Observed frequencies of fish eaten or not eaten by birds according to trematode infection level.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	1	10	37	48
Not eaten by birds	49	35	9	93
Column total	50	45	46	141

TABLE 9.4-2 Expected frequencies of fish eaten and not eaten by birds, according to trematode infection status.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	17.0	15.3	15.7	48
Not eaten by birds	33.0	29.7	30.3	93
Column total	50	45	46	141

37

Expected frequencies shortcut calculation

TABLE 9.4-1 Observed frequencies of fish eaten or not eaten by birds according to trematode infection level.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	1	10	37	48
Not eaten by birds	49	35	9	93
Column total	50	45	46	141

TABLE 9.4-2 Expected frequencies of fish eaten and not eaten by birds, according to trematode infection status.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	17.0	15.3	15.7	48
Not eaten by birds	33.0	29.7	30.3	93
Column total	50	45	46	141

$$Expected[r_i, c_j] = \frac{(row\ i\ total)(column\ j\ total)}{grand\ total}$$

$$Expected[r_1, c_1] = \frac{(48)(50)}{141} = 17.0$$

38

χ^2 contingency statistic

TABLE 9.4-1 Observed frequencies of fish eaten or not eaten by birds according to trematode infection level.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	1	10	37	48
Not eaten by birds	49	35	9	93
Column total	50	45	46	141

TABLE 9.4-2 Expected frequencies of fish eaten and not eaten by birds, according to trematode infection status.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	17.0	15.3	15.7	48
Not eaten by birds	33.0	29.7	30.3	93
Column total	50	45	46	141

$$\chi^2 = \sum_{row=1}^{r} \sum_{col=1}^{c} \frac{[Observed(r,c) - Expected(r,c)]^2}{Expected(r,c)}$$

$r = row$

$c = column$

$$\chi^2 = 69.5$$

39

Degrees of freedom

TABLE 9.4-1 Observed frequencies of fish eaten or not eaten by birds according to trematode infection level.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	1	10	37	48
Not eaten by birds	49	35	9	93
Column total	50	45	46	141

TABLE 9.4-2 Expected frequencies of fish eaten and not eaten by birds, according to trematode infection status.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	17.0	15.3	15.7	48
Not eaten by birds	33.0	29.7	30.3	93
Column total	50	45	46	141

$$df = (r-1)(c-1)$$

$r = row$

$c = column$

$$df = (2-1)(3-1) = 2$$

40

P-value and conclusion

- $\chi_2^2 = 69.5$
- *P*-value based on χ_2^2 distribution: < 10^{-10}
- Below $\alpha = 0.05$, so you reject the null hypothesis of independence between parasite infection and being eaten
- Now go back and look at the data…
- Where are the biggest differences?
  - Uninfected fish eaten far fewer than expected
  - Highly infected fish eaten far higher than expected

TABLE 9.4-1 Observed frequencies of fish eaten or not eaten by birds according to trematode infection level.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	1	10	37	48
Not eaten by birds	49	35	9	93
Column total	50	45	46	141

TABLE 9.4-2 Expected frequencies of fish eaten and not eaten by birds, according to trematode infection status.

	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	17.0	15.3	15.7	48
Not eaten by birds	33.0	29.7	30.3	93
Column total	50	45	46	141

41

Notes

- Skipping section 9.5

42