Unit 2 – Introduction to Probability
Homework #4 (Unit 2 – Introduction to Probability)

SOLUTIONS

1. **These exercises are intended to give you practice in thinking about the real world meanings of some of the measures of association. See unit 2 notes, section 9, *Probabilities in Practice*, especially pp 35-48.**

In introductory epidemiology, one of the study designs that are introduced is the (prospective) **cohort study**. In this type of study involving two groups, the investigator enrolls set (set by design) numbers of participants into each of the two groups that are generically described as "exposed" and "not exposed" and follows them forward to a designated end of the observation period, at which point one or more outcomes are measured.

The following table is from a **cohort study** of Danish men and women that investigated two outcomes, alcohol intake and mortality, in relationship to a number of possible influences: sex, age, body mass index, and smoking. Shown in this table is a cross-tabulation of alcohol intake and death, by sex and level of alcohol intake.

Table 8.2 The distribution of alcohol intake and deaths by sex and level of alcohol intake. Reproduced from *BMJ*, **308**, 302–6, courtesy of BMJ Publishing Group

	Men		Women	
Alcohol intake (beverages a week)*	No of subjects	No (%) of deaths	No of subjects	No (%) of deaths
<1	625	195 (31.2)	2472	394 (15.9)
1–6	1183	252 (21.3)	3079	283 (9.2)
7–13	1825	383 (21.0)	1019	96 (9.4)
14–27	1234	285 (23.1)	543	46 (8.5)
28–41	585	118 (20.2)	72	6 (8.3)
42–69	388	99 (25.5)	29	5 (17.2)
> 69	211	66 (31.3)	20	1 (5.0)
Total	6051	1398 (23.1)	7234	831 (11.5)

*One beverage contains 9–13 g alcohol.

(a) From the information in the table, construct a table with 2 rows and 2 columns. Define your rows by sex and your columns by mortality. What you will have constructed is called a **contingency table**, and specifically, a **2x2 table.**

Some preliminary calculations to get the numbers ….

Men	dead	alive	row total
	195	430	625
	252	931	1183
	383	1442	1825
	285	949	1234
	118	467	585
	99	289	388
	66	145	211
Column total	1398	4653	6051

Women	dead	alive	row total
	394	2078	2472
	283	2796	3079
	96	923	1019
	46	497	543
	6	66	72
	5	24	29
	1	19	20
Column total	831	6403	7234

Answer:

2x2 table

	Dead	Alive	
Men	1398	4653	6051
Women	831	6403	7234
	2229	11056	13285

(b) Next, construct the following contingency table, again with 2 rows and 2 columns.
Define your first row to be persons who consume less than one beverage per week.
Define your second row to be persons who consume more than 69 beverages per week
Define your columns by mortality.

Some preliminary calculations

Men

	Dead	Alive	Row Total
Less than 1 drink/week	**195**	**430**	**625**
More than 69 drinks/week	**66**	**145**	**211**
	261	**575**	**836**

Women

	Dead	Alive	Row Total
Less than 1 drink/week	**394**	**2078**	**2472**
More than 69 drinks/week	**1**	**19**	**20**
	395	**2097**	**2492**

Answer is the sum of the two tables. For example, in row 1 & column 1, 589=195+394:

	Dead	Alive	
Less than 1 drink/week	**589**	**2508**	**3097**
More than 69 drinks/week	**67**	**164**	**231**
	656	**2672**	**3328**

(c) Using the information in your 2x2 table that you constructed in Exercise 1b, calculate the risk of death among persons who consume less than one beverage per week. Then calculate the risk of death among persons who consume more than 69 beverages per week.

	Dead	Alive		Risk of Death =	
Less than 1 drink/week	589	2508	3097	**589/3097 =**	**0.190184049**
More than 69 drinks/week	67	164	231	**67/231=**	**0.29004329**
	656	2672	3328		

(d) In 1-2 sentences, compare the two risk estimates you obtained in Exercise 1c.

The estimated risk of death is approximately 1.5 times greater for persons who drink more than 69 drinks/week (29% risk) relative to those who drink less than 1 drink/week (19% risk).

2. **This question is an elaboration of the thinking that was developed in question 1**.

Another study design that is introduced in introductory epidemiology is the **case-control study**. This study design also calls for the comparison of two groups. Here, however, the investigator enrolls set (again, set by design) numbers of participants, defined by their disease status at the start of the study. **"Cases"** are the enrollees with disease. **"Controls"** are the enrollees who do not have the disease under investigation. The investigation involves looking back in time ("retrospective review") at the histories of all study participants. The goal of this "back in time" look is to see if the cases are different from the controls with respect to their history of some exposure of interest.

The table below is from a **case-control study** that investigated the relationship of occurrences of Down Syndrome (**cases**) to history of exposure to maternal smoking during pregnancy. Shown in the table are some characteristics of the mothers, together with their status with respect to their history of smoking during pregnancy.

Table 8.3 Basic characteristics of mothers in a case-control study of maternal smoking and Down syndrome. Reproduced from *Amer. J. Epid.*, 149, 442–6, courtesy of Oxford University Press

Selected characteristics of Down syndrome cases and birth-matched controls. Washington State, 1984–1994

	Cases (n = 775)		Controls (n = 7750)	
	No.	%	No.	%
Smoking during pregnancy				
Age < 35 years				
Yes	112	20.0	1411	20.2
No	421	75.0	5214	74.6
Unknown	28	5.0	363	5.2
Aged ≥ 35 years				
Yes	15	7.0	108	14.2
No	186	86.9	611	80.2
Unknown	13	6.1	43	5.6

(a) Using the information in the table, construct separate 2x2 contingency tables, one for mothers aged < 35 years and the other for mothers aged $\geq$ 35 years. Define rows by exposure (smoked during pregnancy versus not). Define columns by case status (cases versus controls).

Age < 35

	Case	Control	
Hx smoking during pregnancy	**112**	**1411**	**1523**
Did not smoke during pregnancy	**421**	**5214**	**5635**
	533	**6625**	**7158**

Age $\geq$ 35

	Case	Control	
Hx smoking during pregnancy	**15**	**108**	**123**
Did not smoke during pregnancy	**186**	**611**	**797**
	201	**719**	**920**

(b) For each of the 2x2 tables you constructed in Exercise #2a, calculate two odds:
(i) Odds of smoking during pregnancy among cases
(ii) Odds of smoking during pregnancy among controls

Age < 35

	Case	Control	
Hx smoking during pregnancy	112	1411	1523
Did not smoke during pregnancy	421	5214	5635
	533	6625	7158

	Cases	Controls
Odds of hx smoking =	**112/421=**	**1411/5214=**
	0.266033254	**0.270617568**

Age $\geq$ 35

	Case	Control	
Hx smoking during pregnancy	15	108	123
Did not smoke during pregnancy	186	611	797
	201	719	920

	Cases	Controls
Odds of hx smoking =	**15/186=**	**108/611=**
	0.080645161	**0.176759411**

(c) Using the calculations of odds that you obtained in Exercise #2b, calculate two odds ratios:

(i) Odds Ratio for history of maternal smoking among mothers age < 35 **= 0.98**

(ii) Odds Ratio for history of maternal smoking among mothers age ≥ 35 **= 0.46**

Age < 35

	Case	Control	
Hx smoking during pregnancy	112	1411	1523
Did not smoke during pregnancy	421	5214	5635
	533	6625	7158

	Cases	Controls	**OR=odds of hx(cases)/odds of hx(controls)**
Odds of hx smoking =	112/421=	1411/5214=	**0.26603/0.2706=**
	0.266033254	0.270617568	**0.983059807**

Age ≥ 35

	Case	Control	
Hx smoking during pregnancy	15	108	123
Did not smoke during pregnancy	186	611	797
	201	719	920

	Cases	**Controls**	**OR=odds of hx (cases)/odds of hx (controls)**
Odds of hx smoking =	**15/186=**	**108/611=**	**0.0806/0.1768=**
	0.080645161	**0.176759411**	**0.456242533**

(d) In 1-2 sentences, interpret your results in Exercise 2c.

This case-control study provides no evidence of an adverse association of maternal smoking during pregnancy and Down Syndrome births. Among mothers < 35 years of age, the estimated odds ratio (OR = 0.98) is nearly equal to the null value of 1. Among mothers ≥ 35 years of age, the estimated odds ratio (OR = 0.46) is substantially less than 1.

3. **This question is intended to re-enforce your appreciation of the distinction between the two study designs: prospective cohort versus case-control.**

In 1-2 sentences, why can't you calculate risk in a case-control study?

In a case-control study, participants are not selected on the basis of their exposure to the predictor of interest and then followed for the occurrences of the outcome, which would then permit the estimation of risk. Instead, participants are selected on the basis of their *already* having the outcome or not; indeed, these might even be equal sample sizes. The column totals in your 2x2 table therefore cannot be used to estimate risk of outcome.

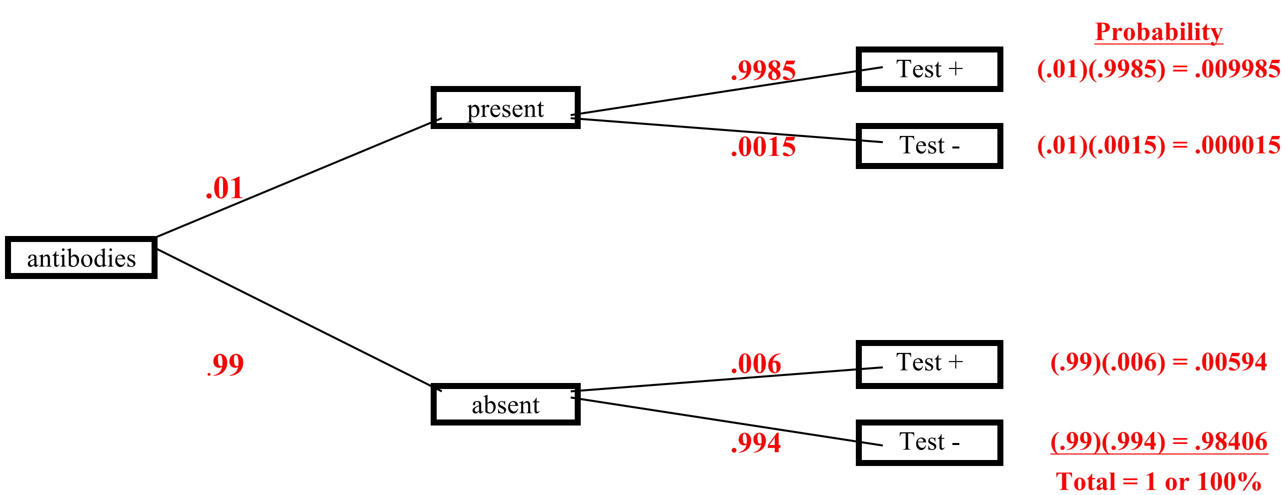
4. **This last question gives you practice thinking about diagnostic tests and the use of Bayes Rule.**

Enzyme immunoassay tests are used to screen blood specimens for the presence of antibodies to HIV, the virus that causes AIDS. The presence of antibodies indicates the presence of the HIV virus. The test is quite accurate but is not always correct. The following table gives the probabilities of positive and negative test results when the blood tested does and does not actually contain antibodies to HIV.

	Test Result	
	Positive (+)	Negative (-)
Antibodies present	0.9985	0.0015
Antibodies absent	0.0060	0.9940

Suppose that 1% of a large population carries antibodies to HIV in their blood.

(a) Draw a tree diagram for selecting a person from this population (outcomes: antibodies present or absent) and for testing his or her blood (outcomes: test positive or negative).



(b) What is the probability that the test is positive for a randomly chosen person for this population? **0.0159, representing a 1.6% chance, approximately.**

The tree shows 4 mutually exclusive outcomes for a person who either has or does not have the antibody and who either tests positive or negative.

Thus, the answer is obtained by summing the probability of the mutually exclusive outcomes that satisfy the event of a positive test.

Pr[test positive] = Pr[antibody and positive test] + Pr[NO antibody and positive test]
= .009985 + .00594
= .015925

(c) What is the probability that a person in this population has the HIV virus, given that he or she tests negative? **0.0000152, representing a 0.0015% chance, approximately.**

Bayes Rule

$$\text{Pr[antibody|test-]} = \frac{\text{Pr[antibody and test-]}}{\text{Pr[test-]}}$$

$$= \frac{\text{pr[antibody]*pr[test-|antibody]}}{\text{pr[antibody]*pr[test-|antibody]+pr[NOantibody]*pr[test-|Noantibody]}}$$

$$= \frac{(.01)(.0015)}{(.01)(.0015)+(.99)(.994)}$$

$$=.0000152$$