EPI 202 Lab 1 Practice Problem Proposed Solutions

1. **In a study of malnutrition and diarrhea in children under age five in Sudanese community, 154 episodes of diarrhea were observed in 38.51 person-years of observation among children who were 75% or less of the expected weight for their age, and 287 episodes of diarrhea were observed in 88.83 person-years of observation among children who were 90% or more of the expected weight for their age (El Samani *et al.*; Am J Epidemiol, 1988).**

   a. Construct an appropriate table to describe the data, and calculate the incidence rates for each of the two groups.

	≤ 75% WFA	≥ 90% WFA	
Cases	a = 154	b = 287	$M_1 = 441$
PY	$N_1 = 38.51$	$N_0 = 88.83$	T = 127.34
$\hat{IR}$	4.00 cases / PY	3.23 cases / PY	

   b. Test the hypothesis that there is no association between malnutrition and the incidence rate of diarrhea. Interpret the numeric results succinctly.

   H_0: The rate of diarrhea does not differ between those children who are less than 75% of their expected weight for age compared to children who are 90% or more of their expected weight for age ($IR_1=IR_0$ or IRR=1 or IRD=0)
   H_A: The rate of diarrhea does differ between those children who are less than 75% of their expected weight for age compared to children who are 90% or more of their expected weight for age ($IR_1 \neq IR_0$ or IRR≠1 or IRD≠0).

   Let X = number of exposed cases = a = 154
   $E(X \mid H_0) = M_1 N_1 / T = 441 * 38.51 / 127.34 = 133.37$
   $Var(X \mid H_0) = M_1 (N_1 / T) \{1-(N_1/T)\} = M_1 N_1 N_0 / T^2$
   $= 441 * 38.51 * 88.83 / 127.34^2 = 93.04$

   Then

$$Z^2 = \frac{[X - E(X \mid H_0)]^2}{Var(X \mid H_0)} = \frac{[154 - 133.37]^2}{93.04} = 4.57$$

   $Pr(X^2 \geq 4.57) = 0.040$

We reject the null at the alpha=0.05 level. We conclude that there is statistically significant evidence for an association between weight for age in children under 5 year (< 75% WFA compared to > 90% WFA) and the incidence rate of diarrhea, assuming no confounding, selection bias, or information bias.

c. Calculate the incidence rate ratio and 95% confidence interval for the association between malnutrition and diarrhea. Interpret the numeric results succinctly.

$$\hat{IRR} = \frac{\hat{IR}_1}{\hat{IR}_0} = \frac{4.00}{3.23} = 1.24$$

Children under age five who are less than 75% of their expected weight for age have an incidence rate of diarrhea that is 24% higher than children who are more than 90% of their expected weight for age, assuming no confounding, selection bias, or information bias.

95% confidence interval for log IRR:

$$\begin{aligned} X \pm 1.96\sqrt{\hat{Var}(X)} &= \log \hat{IRR} \pm 1.96\sqrt{\hat{Var}(\log \hat{IRR})} \\ &= \log \hat{IRR} \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b}} \\ &= \log 1.24 \pm 1.96\sqrt{\frac{1}{154} + \frac{1}{287}} \\ &= (0.017,\ 0.409) \end{aligned}$$

95% confidence interval for IRR:

$e^{(0.17,\ 0.409)} = (1.02,\ 1.51)$

These data are consistent with incidence rate ratios ranging from 1.02 to 1.51 with 95% confidence for the association between malnutrition and diarrhea, assuming no confounding, selection bias, or information bias.

d. **Calculate the incidence rate difference and 95% confidence interval for the association between malnutrition and diarrhea. Interpret the numeric results succinctly.**

$$\hat{IRD} = \hat{IR}_1 - \hat{IR}_0 = 4.00 cases/PY - 3.23 cases/PY = 0.77 cases/PY = 77 \text{ cases/100 PY}$$

These data indicate an incidence rate of diarrhea that is 77 events per 100 person-years higher for children under age five who are less than 75% of their expected weight for age compared with children who are more than 90% of their expected weight for age, assuming no confounding, selection bias, or information bias.

95% confidence interval for IRD:

$$X \pm 1.96\sqrt{\hat{Var}(X)} = \hat{IRD} \pm 1.96\sqrt{\hat{Var}(\hat{IRD})}$$
$$= \hat{IRD} \pm 1.96\sqrt{\frac{a}{N_1^2} + \frac{b}{N_0^2}}$$
$$= 0.77 \pm 1.96\sqrt{\frac{154}{38.51^2} + \frac{287}{88.83^2}}$$
$$= (0.04 \text{cases /PY, } 1.50 \text{ cases/PY})$$

These data are consistent with incidence rate differences ranging from 4 cases/100 person-years to 150 cases/100 person-years with 95% confidence for the association between malnutrition and diarrhea, assuming no confounding, selection bias, or information bias.

2. A cigarette smoking history was obtained from 6,690 Japanese-American men examined from 1965 through 1968. Over 22-years of follow-up, 37 incident cases of oral or bladder cancer were observed among the 2,344 never smokers, and 165 incident cases of oral or bladder cancer were observed among the 4,346 past and current smokers (Chyou PH, Nomura AMY, Stemmermann GN. Am J Public Health 1992;83:37-40). Assume no loss to follow-up and no competing causes of death in all subsequent questions.

a. Construct an appropriate table to describe the data, and calculate the cumulative incidence of oral or bladder cancer for each of the two groups.

	Ever smoked	Never smoked	Total
Cases	a = 165	b = 37	M_1 = 202
Non-cases	c = 4181	d = 2307	M_0 = 6488
Total	N_1 = 4346	N_0 = 2344	T = 6690
$\hat{C}$	165 / 4346 = 0.038	37 / 2344 = 0.016	
	(over the 22-year study period)		

b. Test the hypothesis that smoking has no association with the 22-year cumulative incidence of oral or bladder cancer. Interpret the numerical results succinctly.

H_0: Cigarette smoking is not associated with the 22-year cumulative incidence of oral or bladder cancer, i.e. $C_1 = C_0$
H_A: Cigarette smoking is associated with the 22-year cumulative incidence of oral or bladder cancer, i.e. $C_1 \neq C_0$

Let X = number of exposed cases = a = 165
E (X | H_0) = N_1M_1/T = (4346*202)/6690 = 131.225
Var (X | H_0) = $M_1M_0N_1N_0/T^3$ = (202*6488*4346*2344)/6690^3 = 44.589

$$Z^2 = \frac{[X - E(X|H_0)]^2}{Var(X|H_0)} = \frac{(165 - 131.225)^2}{44.589} = 25.6$$

$\Pr[X^2 > 25.6] < 0.0001$

An alternative approach:

Let $X = \hat{CID} = \hat{C}_1 - \hat{C}_0 = \frac{165}{4346} - \frac{37}{2344} = 0.0222$ over the 22-year study period

E (X | H_0) = 0

$$Var(X \mid H_0) = \bar{C}(1-\bar{C})\left(\frac{1}{N_1} + \frac{1}{N_0}\right), \text{ where } \bar{C} = \frac{M_1}{T} = \frac{202}{6690}\left(1 - \frac{202}{6690}\right)\left(\frac{1}{4346} + \frac{1}{2344}\right)$$

= 0.00001923

Then

$$Z^2 = \frac{[X - E(X \mid H_0)]^2}{Var(X \mid H_0)} = \frac{[0.0222 - 0]^2}{0.00001923} = 25.6$$

$\Pr[X^2 > 25.6] < 0.0001$

We reject the null at the alpha=0.05 level. There is statistically significant evidence of an association between cigarette smoking and the cumulative incidence of oral or bladder cancer observed in these data over the 22-year study period, assuming no confounding, selection bias, or information bias. (Notice that the hypothesis test gives no indication of the magnitude or direction of the association observed.)

c. Calculate the cumulative incidence ratio for oral or bladder cancer from cigarette smoking, and the relevant 95% confidence limits. Interpret the numerical results succinctly.

$$\hat{CIR} = \frac{\hat{C}_1}{\hat{C}_0} = \frac{165/4346}{37/2344} = 2.41 \text{ over the 22-year study period}$$

Over the 22-year study period, ever smokers have a 2.4-fold higher cumulative incidence of oral or bladder cancer compared to never smokers, assuming no confounding, selection bias, or information bias.

95% confidence interval for log CIR:

$$X \pm 1.96\sqrt{\hat{Var}(X)} = \log \hat{CIR} \pm 1.96\sqrt{\hat{Var}\left(\log \hat{CIR}\right)}$$

$$= \log \hat{CIR} \pm 1.96\sqrt{\frac{N_1 - a}{aN_1} + \frac{N_0 - b}{bN_0}}$$

$$= \log 2.41 \pm 1.96\sqrt{\frac{4346 - 165}{165 * 4346} + \frac{2344 - 37}{37 * 2344}}$$

$$= \log 2.41 \pm 1.96\sqrt{0.0324}$$

= (0.527, 1.233)

95% confidence interval for CIR:

$e^{(0.527,\ 1.233)}$ = (1.69, 3.43) over the 22-year study period

With 95% confidence, these data are consistent with cumulative incidence ratios ranging from 1.69 to 3.43 over the 22-year study period for the association between cigarette smoking and oral or bladder cancer, assuming no confounding, selection bias, or information bias.

d. Calculate the cumulative incidence difference for oral or bladder cancer from cigarette smoking, and the relevant 95% confidence limits. Interpret the numerical results succinctly.

$$\hat{CID} = \hat{C}_1 - \hat{C}_0 = \frac{165}{4346} - \frac{37}{2344} = 0.0222$$ over the 22-year study period

Over the 22-year study period, there is a 2.2% excess cumulative incidence of oral or bladder cancer in ever smokers compared to never smokers, assuming no confounding, selection bias, or information bias.

95% confidence interval for CID:

$$X \pm 1.96\sqrt{\hat{Var}(X)} = \hat{CID} \pm 1.96\sqrt{\hat{Var}(\hat{CID})}$$

$$= \hat{CID} \pm 1.96\sqrt{\frac{a(N_1 - a)}{N_1^3} + \frac{b(N_0 - b)}{N_0^3}}$$

$$= 0.0222 \pm 1.96\sqrt{\frac{165(4346 - 165)}{4346^3} + \frac{37(2344 - 37)}{2344^3}}$$

$$= 0.0222 \pm 1.96\sqrt{0.0000150}$$

= (0.014, 0.030) over the 22-year study period

With 95% confidence, these data are consistent with cumulative incidence differences ranging from 0.014 to 0.030 over the 22-year study period for the association between cigarette smoking and oral or bladder cancer, assuming no confounding, selection bias, or information bias.

3. Please get out your computer & download the example dataset, 'evans_example_dat.csv' from the course website. Once you have downloaded it, use the example code in the documentation for your preferred statistical package to import the data. Documentation is available for SAS, STATA, and R on the course webpage in the 'Statistical Software for Tabular Analysis' module.

If your file is saved at the following pathway: 'P:\My Documents\evans_example_dat.csv'

How to import a file to SAS and create a dataset named 'dat' in the Work folder (code is not case sensitive):

```
proc import out=WORK.dat datafile='P:\My Documents\evans_example_dat.csv' dbms=csv
replace;
getnames=yes;
run;
```

How to import a file to STATA (code is case sensitive):

```
import delimited "P:\My Documents\evans_example_dat.csv"
```

How to import a file to R and create a dataset called 'evansData' (code is case sensitive):

```
#Set your working directory to the location where you saved the data file
setwd("'P:/My Documents")
#Read in the data file, use the header=T option to tell R the first row contains the variable names
evansData<-read.csv("evans_example_dat.csv", header=T)
```