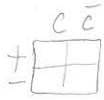

HW 3

1/ In a NY Times article (Feb 17, 1999) about the PSA blood test for detecting prostate cancer, of the men who had the disease, the test fails to detect prostate cancer in 1 in 4; of those without the cancer, 2/3 received false positives. Let $C/\bar{C}$ denote having/not having prostate cancer and let +/− denote a positive/negative result.

a) Which is true: $P(-|C) = \frac{1}{4}$ or $P(C|-) = \frac{1}{4}$? $\boxed{P(-|C) = \frac{1}{4}}$

$P(\bar{C}|+) = \frac{2}{3}$ or $P(+|\bar{C}) = \frac{2}{3}$? $\boxed{P(\bar{C}|+) = \frac{2}{3}}$

b) What is the sensitivity of the test?

Sensitivity $= P(+|C) \boxed{= \frac{3}{4}}$

c) Of men who take the PCA test, suppos $P(C) = .01$. Find the cell probabilities in the 2×2 table for the joint distribution that cross classifies Y = diagnosis (+, −) with X = true disease status $(C, \bar{C})$.

$P(-|C) = .25 \rightarrow .25 \cdot .01 = .0025$

$P(\bar{C}|+) = .6 = \frac{P(\bar{C})}{P(+)} = \frac{.99}{P(+)} \rightarrow P(+) = \frac{P(\bar{C})}{P(\bar{C}|+)} = \frac{.99}{.6}$

X	Y: +	Y: −	
C	.0075	.0025	.01
$\bar{C}$	.0150	.9750	.99
	.0225	.9775	1

d) Using (c), find the marginal distribution for the diagnosis. (above)

e) Using (c) and (d), find $P(C|+)$, and interpret

$P(C|+) = \frac{.0075}{.0225} \boxed{= \frac{1}{3}}$

Approximately $\frac{1}{3}$ men who test positive for prostate cancer will actually have the disease

2] For diagnostic testing, let X = true status (1 = disease, 2 = no disease) and Y = diagnosis (1 = positive 2 = negative). Let $\pi_i = P(Y=1 \mid X=i)$, $i = 1, 2$

a) Explain why sensitivity $= \pi_1$ and specificity $= 1 - \pi_2$

sensitivity $= \pi_1$ because when $Y=1$, $X=1$ we have detected a true positive.

specificity $= 1-\pi_2$ because the complement of π_2 is the case when true negatives are detected.

b) Let γ denote the probability the subject has the disease. Given that the diagnosis is positive, use Baye's theorem to show that the probability the subject truly has the disease is: $\dfrac{\pi_1 \gamma}{\pi_1 \gamma + \pi_2 (1-\gamma)}$

This is Bayes' Theorem rewritten to correspond to our table format.

π_1	π_2
$1-\pi_1$	$1-\pi_2$

$\pi_1 = P(C \mid +)$ and $\gamma = P(C)$

$\pi_2 = P(\bar{C} \mid +)$ $\quad 1-\gamma = P(\bar{C})$

c) For mammograms for detecting breast cancer, suppose $\gamma = .01$, sensitivity $= .86$, and specificity $= .88$. Given a positive test result, find the probability that a woman truly has breast cancer.

$$P(C) = \frac{(.86)(.01)}{(.86)(.01) + (.88)(1-.01)} = \frac{(.86)(.01)}{(.01)(.86) + (.88)(.99)} \boxed{= .009775 \approx .977\%}$$

d) To better understand the answer in (c), find the joint probabilities for the 2x2 cross classification of X and Y. Discuss their relative sizes in the two cells that refer to a positive result.

		Y = C	Y = $\bar{C}$
X	+	.86	.12
	−	.14	.88
		1	1

π_{11} is much larger than π_{12}, which is good because we would not expect for a large number of "successes" to appear if the patient does not have the disease. Otherwise our test would not be valuable because it would indicate that everyone has the disease, whether or not they do.

3] According to UN figures, the annual gun homocide rate is 62.4 per one million residents in the U.S. and 1.3 per one million in the U.K

a) Compare the proportion of annual gun homocides using:

i) Difference of proportions

$\pi_1 - \pi_2 = \frac{62.4}{1,000,000} - \frac{1.3}{1,000,000}$ $= \frac{61.1}{1,000,000}$

ii) Relative Risk

$RR = \frac{\pi_1}{\pi_2} = \frac{\frac{62.4}{1,000,000}}{\frac{1.3}{1,000,000}} = \frac{62.4}{1.3}$ $= 48$

b) When both proportions are very close to 0, as here, which measurement is more useful for describing the strength of association? Why?

Relative Risk because (1) the resut is on a larger, more intuitive scale (everyday numbers instead of near zero ones), and (2) because it relates the proportions in terms of scale, rather than absolute difference.

4] A newspaper article preceding the 1994 World Cup semifinal match between Italy and Bulgaria stated that "Italy is favored 10-11 to beat Bulgaria, which is rated 10-3 to reach the final." Suppose this means that the odds Italy wins are 11/10 and that Bulgaria wins are 3/10. Find the probability each team wins and comment.

$P(\text{Italy wins}) = \frac{11/10}{11/10 + 1}$ $= .52381$

$P(\text{Bulgaria wins}) = \frac{3/10}{3/10 + 1}$ $= .23077$

These outcomes are interesting because they would suggest that there is an outcome other than either of these two countries winning the semifinal game.

$.52 + .23 \neq 1$

5] Consider the following two studies reported in the NY Times:

a) A British study reported (Dec. 3, 1998) that, of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer." Is 1.7 a relative risk, or an odds ratio?

A Relative Risk

* This is a relative risk because it compares two probabilities of two different events. An odds ratio compares the probability of success of one event to the probability of a failure of that event.

b) A National Cancer Institute study about tamoxifen and breast cancer reported (April 7, 1998) that the women taking the drug were 45% less likely to experience invasive breast cancer compared with women taking the placebo. Find the relative risk for:

i) those taking tamoxifen compared to those taking the placebo.

$\pi_1 = (1-.45)\pi_2 \rightarrow \frac{\pi_1}{\pi_2} = \frac{(.55\,\pi_2)}{\pi_2}$ $= .55$

ii) those taking the placebo and those taking tamoxifen.

$\pi_2 = .55\,\pi_1 \rightarrow \frac{\pi_1}{\pi_2} = \frac{\pi_1}{.55\pi_2} = \frac{1}{.55}$ $= 1.81$

6] In the U.S., the estimated annual probability that a woman over the age of 35 dies of lung cancer equals .001304 for current smokers and .000121 for non-smokers [M. Pagano and K. Gauvreau, Principles of Biostatistics, Belmont, CA: Duxbury Press (1993), p. 134].

a) Calculate and interpret the difference of proportions and the relative risk. Which is more informative for these data? Why?

$DoP = (.001304) - (.000121)$ $= .00118.$

$RR = \frac{(.001304)}{(.000121)}$ $= 10.78$

The RR is more informative because it is a more "real" number (more easily understood) than the DoP. It puts the information on a scale that is somewhat independent of the specific probabilities.

b) Calculate and interpret the odds ratio. Explain why the RR and OR take similar values.

$odds_{smoker} = \frac{(.001304)}{1-(.001304)} \approx .001306$ $\quad odds_{non\text{-}smoker} = \frac{(.000121)}{1-(.000121)} \approx .000121$

$\hat{\theta} = \frac{(.001306)}{(.000121)}$ $= 10.79$

The OR and RR are similar because the probability of dying of cancer is near 0, making the odds of each event very near the probability due to the near 1 divisor in the odds formula.

7] For adults who sailed on the Titanic, the odds ratio between gender (male, female) and survival (yes, no) was 11.4.

a) What is wrong with the interpretation: "The probability for survival for females was 11.4 times that for males"? Give the correct interpretation.

That interpretation would be for if the RR, the correct interpretation would be that the odds of a female to survive are 11.4 times the odds of a male surviving.

b) The odds of survival for females equaled 2.9. For each gender, find the proportion of survivors.

$OR = 1.4$ $\quad odds_{male} = \frac{odds_{female}}{OR} = \frac{(2.9)}{(11.4)} = .254$

$\pi_{females} = \frac{(2.9)}{1+(2.9)} = .744$ $\qquad \pi_{males} = \frac{(.254)}{1+(.254)} = .203$

c) Find the value of R in the interpretation, "The probability of survival for females was R times that for males."

$R = \frac{\pi_f}{\pi_m} = \frac{(.744)}{(.203)} = 3.673$

8] A research study estimated that under a ceartain condition, the probability a subject would be referred for heart catheterization was .906 for whites and .847 for blacks.

a) A press release about the study stated the odds referral for cardiac catheterization are 60% of odds for whites. Explain how they obtained 60% (more accurately, 57%).

This computation was done by correctly calculating the odds ratio

b) An AP story that described the study stated "Doctors were only 60% as likely to order cardiac catheterization for blacks as for whites." What is wrong with this interpretation? Give the correct percentage for this interpretation. (In stating results to the general public, it is better to use RR than OR, it is simpler to understand.)

This interpretation is using an OR in place of a RR. The true RR is:

$RR = \frac{p_{black}}{p_{white}} = \frac{(.847)}{(.906)} = .935 \rightarrow 93.5\%$

9] An estimated odds ratio for adult females between the presence of squamous cell carcinoma (yes, no) and smoking behavior (smoker, nonsmoker) equals 11.7 when the smoker category consists of subjects whose smoking level s is $0 < s < 20$ cigarettes per day; It is 26.1 for smokers with $s \geq 20$ cigarettes per day. Show that the estimated odds ratio between carcinoma and smoking level ($s \geq 20$, $0 < s < 20$) equals $\frac{26.1}{11.7} = 2.2$.

$$OR = \frac{\text{odds}(s \geq 20)}{\text{odds}(0 < s < 20)} = \frac{(26.1)}{(11.7)} = 2.2$$

10] Data posted at the FBI website stated that of all blacks slain in 2005, 91% were slain by blacks, and of all whites slain in 2005, 83% were slain by whites. Let Y denote race of victim and X denote race of murderer.

		Victim B	Victim W
Murderer	B	.91	.17
	W	.09	.83

	W	B
W	.83	.09
B	.17	.91

a) Which conditional distribution do these statistics refer to, Y given X or X given Y?

Y given X

b) Calculate the odds ratio between X and Y and interpret.

$\hat{\theta} = \frac{(.91)(.83)}{(.09)(.17)} = 49.366$ This OR indicates that the odds of blacks being slain by blacks is almost 50 times that of whites being slain by blacks.

c) Given that the murderer was white, can you estimate the probability that the victim was white? What additional information would you need to do this? (Hint: How could you use Bayes' Theorem?).

No. If we were given the probability that the murderer was a white person, then using Bayes' theorem, using race of the murderer as the "test outcome (positive or negative) and race of the victim as the "disease" (white, not white) we could find the true probability that the victim was white.

* Assuming only 2 races were involved in the data

12) A statistical analysis that combines information from several studies is called a meta analysis. A meta analysis compared aspirin and placebo on incidence of heart attack and of stroke, separately for men and for women. For the Women's Health Study, heart attacks were reported for 198 of 19,934 taking aspirin and for 193 of 19,942 taking placebo.

a) Construct the 2x2 table that classifies the treatment (aspirin, placebo) with whether a heart attack was reported (yes, no).

Outcome \ Treatment	aspirin	placebo	
heart attack	198	193	391
no heart attack	19,736	19,749	39,485
	19,934	19,942	

b) Estimate the odds ratio. Interpret.

$$\text{odds}_A = \frac{\left(\frac{198}{19,934}\right)}{1-\left(\frac{198}{19,934}\right)} = .01003 \qquad \text{odds}_P = \frac{\left(\frac{193}{19,749}\right)}{1-\left(\frac{193}{19,749}\right)} = .00987$$

$$\hat{\theta}_{A/P} = \frac{(.01003)}{(.00987)} = 1.01655$$

The odds of a woman having a heart attack appear to be slightly raised by taking aspirin.

c) Find a 95% confidence interval for the population odds ratio for women. Interpret. (As of 2006, results suggest that for women, aspirin was helpful for reducing risk of stroke, but not necessarily risk of heart attack.)

$$\ln \hat{\theta} \pm z_{\alpha/2}(SE) = \ln(1.01655) \pm (1.96)\sqrt{\frac{1}{(198)} + \frac{1}{(19,934)} + \frac{1}{(193)} + \frac{1}{(19,942)}}$$

$$= \ln(1.01655) \pm 1.96\sqrt{.01033} \rightarrow (-0.18279,\ 0.21562)$$

$$\rightarrow (e^{-.18279},\ e^{.21562}) = (.83294,\ 1.24063)$$

It would not appear that it can be concluded that aspirin has any direct effect on whether or not a woman will have a heart attack.

13] Refer to table 2.1 about beleif in the afterlife.

Table 2.1 Cross Classification of Beleif in Afterlife by Gender

	Beleif in Afterlife		
Gender	Yes	No or Undecided	Total
F	509	116	625
M	398	104	502

$n = 1127$

a) Construct a 90% confidence interval for the difference of proportions and interpret.

$\hat{p}_F = \frac{509}{625} = .8144 \quad \hat{p}_M = \frac{398}{502} = .79283 \quad DoP = (.8144) - (.79283) = .02157$

$CI: DoP \pm z_{\alpha/2} SE = DoP \pm 1.645 \sqrt{\frac{\hat{p}_F(1-\hat{p}_F)}{n_1} + \frac{\hat{p}_M(1-\hat{p}_M)}{n_2}} = (.02157) \pm 1.645 \sqrt{\frac{(.8144)(.1856)}{(625)} + \frac{(.79283)(.207}{(502)}}$

$= .02157 \pm .03924 = (-.01767, .06081)$

It is inconclusive at the 90% confidence level whether women are more, or less likely to beleive in an afterlife.

b) Construct a 90% confidence interval for the odds ratio and interpret.

$odds_F = \frac{.8144}{1-.8144} = 4.38793 \quad odds_M = \frac{.79283}{1-.79283} = 3.82695 \quad \hat{\theta} = \frac{(4.38793)}{(3.82695)} = 1.14659$

$CI: \ln(1.14659) \pm 1.645 \sqrt{\frac{1}{509} + \frac{1}{398} + \frac{1}{116} + \frac{1}{104}} = (-.11113, .38471)$

$\rightarrow (e^{-.11113}, e^{.38471}) = (.89482, 1.46918)$

Again, we cannot conclude whether or not the odds are in favor of females beleiving in the afterlife more often than men.

c) Conduct a test of statistical independence. Report the p-value and interpret.

$\hat{\mu} = n p_{i+} p_{+j} = n \left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+} n_{+j}}{n}, \quad \hat{\mu}_{11} = \frac{(625)(907)}{(1127)} = 502.995$

$\hat{\mu}_{12} = \frac{(625)(220)}{(1127)} = 122.005 \quad \hat{\mu}_{21} = \frac{(502)(907)}{(1127)} = 404.005 \quad \hat{\mu}_{22} = \frac{(502)(220)}{(1127)} = 97.995$

$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} = \frac{(509-502.995)^2}{502.995} + \frac{(116-122.005)^2}{122.005} + \frac{(398-404.005)^2}{404.005} + \frac{(104-97.995)^2}{97.995} = .8245$

$df = (2-1)(2-1) = 1 \rightarrow p > .250$

This P-value suggests that the difference between males and females in beleif in afterlife is not significant