

**PubHlth 540 - Introductory Biostatistics  
Fall 2008  
Examination I**

**SOLUTIONS**

**1. (10 points total)**

The Honolulu Heart Study, conducted in the 1970's, measured lots of things on lots of study participants. The following are 10 of the study variables. For each, please tell me two things: (1) whether the variable measured was qualitative or quantitative. If qualitative, indicate (2) whether the variable measured was nominal or ordinal. If quantitative, indicate (2) whether the variable measured was discrete or continuous

		Qualitative or Quantitative	If qualitative: nominal or ordinal If quantitative: discrete or continuous
<b>1 point</b>	<u>Education level:</u> 1=none 2=primary 3=intermediate 4=senior high 5=technical school 6=university	<b>Qualitative</b>	<b>Ordinal</b>
<b>1 point</b>	<u>Weight:</u> kilograms	<b>Quantitative</b>	<b>Continuous</b>
<b>1 point</b>	<u>Height:</u> centimeters	<b>Quantitative</b>	<b>Continuous</b>
<b>1 point</b>	<u>Age:</u> Years	<b>Quantitative</b>	<b>Continuous</b>
<b>1 point</b>	<u>Smoking status:</u> 0=no 1=yes	<b>Qualitative</b>	<b>Nominal</b>
<b>1 point</b>	<u>Physical activity:</u> 1=mostly sitting 2= moderate 3=active	<b>Qualitative</b>	<b>Ordinal</b>
<b>1 point</b>	<u>Blood glucose:</u> Milligrams percent	<b>Quantitative</b>	<b>Continuous</b>
<b>1 point</b>	<u>Serum cholesterol:</u> Milligrams/deciliter	<b>Quantitative</b>	<b>Continuous</b>
<b>1 point</b>	<u>Systolic blood pressure</u> Millimeters mercury	<b>Quantitative</b>	<b>Continuous</b>
<b>1 point</b>	<u>Ponderal index:</u> $\text{Height}^3/\text{weight}$	<b>Quantitative</b>	<b>Continuous</b>

**2. (10 points total)**

The following is a table showing the frequency and cumulative frequency distribution of the survival times of 347 patients diagnosed with cancer.

Survival Time (years)	Frequency	Cumulative Frequency
<1	62	62
1-2	45	107
2-3	38	145
3-4	28	173
4-5	25	198
5-6	10	208
6-7	14	222
7-8	11	233
8-9	9	242
9-10	8	250
10-11	8	258
11-12	8	266
12-13	9	275
13-14	5	280
14-15	2	282
15-16	3	285
16-17	4	289
17-18	7	296
18-19	1	297
19-20	3	300
At least 20	47	347
<b>Total</b>	<b>347</b>	

- (a) **(3 points)** What is the **modal** survival time? **<1 years. The mode is the interval of survival time with the greatest frequency.**
- (b) **(3 points)** Estimate the **median** survival time. **4-5 years. The median survival time separates the lower 50% from the upper 50% of the individual survival times. For an ordered sample size of 347, this would be the  $(347 + 1)/2^{\text{th}} = 174^{\text{th}}$  ordered observation. The endpoint of the interval 3-4 years corresponds to a cumulative frequency of 173. Therefore, the next interval, 4-5 years, ensures the 50-50 split.**
- (c) **(2 points)** Estimate the smallest value of the **mean** survival time. **I will accept a variety of solutions here. One solution might be to do a weighted mean using the midpoints of each interval for all but the last interval and using the value 20 years for the last interval. Another solution might be similar but using instead the minimum of each interval. I used excel here:**

## Using midpoint

<u>interval</u>	<u>midpoint</u>	<u>freq</u>	<u>midpoint*freq</u>
<1	0.5	62	31
1-2	1.5	45	67.5
2-3	2.5	38	95
3-4	3.5	28	98
4-5	4.5	25	112.5
5-6	5.5	10	55
6-7	6.5	14	91
7-8	7.5	11	82.5
8-9	8.5	9	76.5
9-10	9.5	8	76
10-11	10.5	8	84
11-12	11.5	8	92
12-13	12.5	9	112.5
13-14	13.5	5	67.5
14-15	14.5	2	29
15-16	15.5	3	46.5
16-17	16.5	4	66
17-18	17.5	7	122.5
18-19	18.5	1	18.5
19-20	19.5	3	58.5
at least 20	20	47	940

6.979827089 weighted mean

## Using lower bound

<u>interval</u>	<u>midpoint</u>	<u>freq</u>	<u>midpoint*freq</u>
<1	0.5	62	31
1-2	1	45	45
2-3	2	38	76
3-4	3	28	84
4-5	4	25	100
5-6	5	10	50
6-7	6	14	84
7-8	7	11	77
8-9	8	9	72
9-10	9	8	72
10-11	10	8	80
11-12	11	8	88
12-13	12	9	108
13-14	13	5	65
14-15	14	2	28
15-16	15	3	45
16-17	16	4	64

17-18	17	7	119
18-19	18	1	18
19-20	19	3	57
at least 20	20	47	940

**6.636887608 weighted mean**

- (d) **(2 points)** In your opinion, are these data symmetric, positively skewed, or negatively skewed? Explain your answer. **These data are positively skewed, since mean > median. Note also that mean > median > mode!**

**3. (5 points total)**

Consider the relationship between the standard deviation and the standard error. Suppose it is known that the standard deviation is 3. How large a sample  $n$  should be taken for the standard error of the mean to have a value of 0.5? **36.**

$$SE(\bar{X}) = \frac{SD}{\sqrt{n}} \Rightarrow n = \left( \frac{SD}{SE} \right)^2 \Rightarrow n = \left( \frac{3}{0.5} \right)^2 = 36$$

## 4. (10 points total)

Consider the following cross-tabulation of 101 individuals by their smoking status and systolic blood pressure. Shown are counts of individuals. For example, the entry “10” in the first row says that there are 10 non-smokers in this sample with systolic blood pressure between 90 and 109 mm Hg.

Systolic blood pressure, mm Hg	Non-Smokers	Smokers
90-109	10	5
110-129	24	15
130-149	18	10
150-169	9	3
170-189	2	2
190-209	0	3

Consider this distribution as a population and, as such, a universe of possibilities from which simple probabilities can be computed.

Define two events “A” and “B” as follows:

A = smoker

B = systolic blood pressure of 170 or greater.

(a) (2 points) Find Probability [ A ]

$$p(A) = (\text{number of smokers})/(\text{total number of individuals}) = 38/101 = 0.38$$

(b) (2 points) Find Probability [ B].

$$p(B) = (\text{number of individuals with SBP} \geq 170)/(\text{total number of individuals}) \\ p(B) = 7/101 = 0.07$$

(c) (6 points) Are A and B independent? Explain.

For the two probabilities to be independent the following must hold:

$$P(A|B) = P(A)$$

$$p(A|B) = (\text{number of smoker individuals with SBP} \geq 170)/(\text{number individuals with SBP} \geq 170) \\ = 5/7 \\ = 0.714$$

$$p(A) = (\text{number of smokers})/(\text{total number of individuals}) \\ = 38/101 \\ = 0.38$$

Thus,  $P(A|B) \neq P(A)$  and the events are not independent.

**5. (15 points total)**

It's not just an exercise in algebra that there is sometimes interest in re-location and re-scaling. One example of re-location and re-scaling is the conversion of temperature information expressed in degrees centigrade to degrees Fahrenheit. A second example, closer to public health, is the use of international units IU. Weight reported in kilograms or pounds and height reported in centimeters or inches are other examples.

Suppose that the sample mean and sample standard deviation of  $n=50$  values of a random variable  $X$  are  $\bar{X} = 100$  and  $s = 15$ . (note – this is close to the values of the population mean and standard deviation of IQ).

(a) **(5 points)** A constant  $b=5$  is added to each observation. What are the new values of the

- 1 point.** Sample mean =  $100+5 = 105$ ,
- 1 point** Sample standard deviation =  $15$
- 3 points** Sample variance =  $225$

(b) **(5 points)** Next, consider instead that each observation is multiplied by a constant  $a=5$ . What are the new values of the

- 1 point.** Sample mean =  $(5)(100) = 500$ ,
- 1 point** Sample standard deviation =  $(5)(15) = 75$
- 3 points** Sample variance =  $(5^2)(15^2) = 5625$

(c) **(5 points)** Now suppose that the sample mean and sample standard deviation of 100 measurements of temperature in centigrade are  $\bar{X}$  and  $s$ . What is the expression for the same sample mean and sample standard deviation when the temperature values are converted to Fahrenheit?

*Hint: [Degrees Fahrenheit] = (1.8)[Degrees centigrade] + 32*

$$\text{Sample mean in fahrenheit} = (1.8)(\bar{X}_{\text{centigrade}}) + 32$$

$$\text{Sample standard deviation in Fahrenheit} = (1.8)(S_{\text{centigrade}})$$

**6. (10 points total)**

- (a) **(5 points)** The following is a relative frequency table summarizing the score earned on a PubHlth 540 exam. What is the average score?

Exam I Score	Relative Frequency
100	.40
88	.30
81	.20
76	.10

**Average score= 90.2**

$$= \sum_{i=1}^5 [\text{Exam I Score}]_i [\text{Relative Frequency}]_i = (100)(0.4) + (88)(0.3) + (81)(0.2) + (76)(0.1) =$$

$$= 40+26.4+16.2+7.6=90.2$$

- (b) **(5 points)** As the Fall 2008 election draws near, we find ourselves on the brink of a disastrous recession. During recessions, average earnings typically go up. Explain how this can be.

**If unemployed people are excluded from the average earnings calculation (this is often done) and if some people lose their jobs during a recession (this often happens), and if the layoffs are disproportionately often the low income salary earners (this often happens), then the new average will be of employed persons with higher salaries.**

7. (10 points total)

Consider the following data on a new screening test for diabetes

	Diabetes	NO diabetes	Total
Test +	59	48	107
Test -	11	462	473
	70	510	580

(a) (4 points) Using the information in the table, compute the following:

- 1 point. Sensitivity =  $59/70=0.843$ ,
- 1 point. Specificity =  $462/510=0.906$
- 1 point. Predictive value positive =  $59/107=0.551$
- 1 point. Predictive value negative =  $462/473=0.977$

(b) (4 points) Using the values of sensitivity and specificity from part “a” and *not using* your values of predictive value positive from part “a”, use your understanding of *Bayes rule* to complete the following table.

The solution is exactly like the example on page 30 of the topic 2 notes.

$$\begin{aligned}
 \text{Predictive Value + Test} &= \frac{\text{Pr}(+\text{test}|\text{disease})\text{Pr}(\text{disease})}{\text{Pr}(+\text{test}|\text{disease})\text{Pr}(\text{disease}) + \text{Pr}(+\text{test}|\text{NOdisease})\text{Pr}(\text{NOdisease})} \\
 &= \frac{(\text{sensitivity})(\text{prevalence})}{(\text{sensitivity})(\text{prevalence}) + (1-\text{specificity})(1-\text{prevalence})}
 \end{aligned}$$

Prevalence Diabetes	Predictive Value + Test
.01	<b>.0831</b>
.05	<b>.3207</b>
.10	<b>.4991</b>
.20	<b>.6916</b>
.30	<b>.7935</b>
.40	<b>.8567</b>
.50	<b>.8997</b>
.60	<b>.9308</b>
.70	<b>.9544</b>
.80	<b>.9729</b>
.90	<b>.9878</b>
.95	<b>.9942</b>
.99	<b>.9989</b>

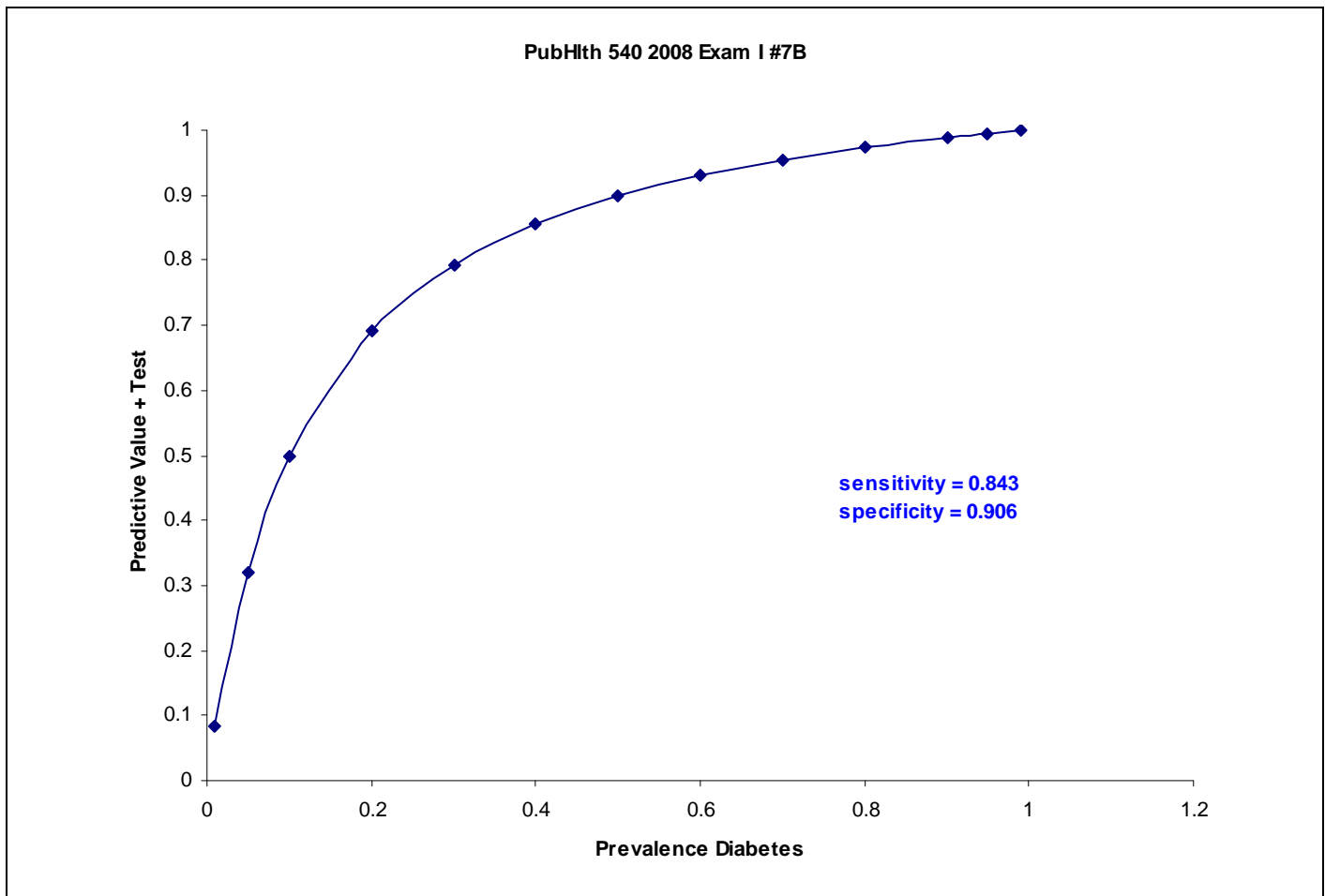
Here is the solution in excel so you can see all the work:

In what follows, I used sensitivity = .843, specificity = .906, and (1-specificity)=.094

prevalence sensitivity (prev)(sens) (1-specif) (1-prev) (1-spec)(1-prev) numerator denom Pred value +

0.01	0.843	0.00843	0.094	0.99	0.09306	0.00843	0.10149	0.083062371
0.05	0.843	0.04215	0.094	0.95	0.0893	0.04215	0.13145	0.320654241
0.1	0.843	0.0843	0.094	0.9	0.0846	0.0843	0.1689	0.499111901
0.2	0.843	0.1686	0.094	0.8	0.0752	0.1686	0.2438	0.691550451
0.3	0.843	0.2529	0.094	0.7	0.0658	0.2529	0.3187	0.793536241
0.4	0.843	0.3372	0.094	0.6	0.0564	0.3372	0.3936	0.856707317
0.5	0.843	0.4215	0.094	0.5	0.047	0.4215	0.4685	0.899679829
0.6	0.843	0.5058	0.094	0.4	0.0376	0.5058	0.5434	0.930806036
0.7	0.843	0.5901	0.094	0.3	0.0282	0.5901	0.6183	0.954391072
0.8	0.843	0.6744	0.094	0.2	0.0188	0.6744	0.6932	0.9728794
0.9	0.843	0.7587	0.094	0.1	0.0094	0.7587	0.7681	0.98776201
0.95	0.843	0.80085	0.094	0.05	0.0047	0.80085	0.80555	0.994165477
0.99	0.843	0.83457	0.094	0.01	0.00094	0.83457	0.83551	0.998874939

If you used a calculator on the web, that's fine, too. Just for fun, I graphed it in excel, too. We see clearly that the predictive value positive of a given test changes depending on the prevalence of disease.



(c)  
**(2 points)** True or False.

**1 point. True** The sensitivity of a diagnostic test is independent of the probability of disease

**1 point. False** The predictive value of a diagnostic test is independent of the probability of disease.

**8. (10 points total)**

A scientist has just been funded by the National Institutes of Health to conduct a very expensive study. As it happens, the budget has been cut. Therefore, the choice of laboratory for analysis of sample data is being re-visited.

Suppose the choice is between laboratory “A” and “B”. Laboratory “A” has a standard deviation of one milligram and each sample analysis costs \$3. Laboratory “B” has a standard deviation of two milligrams but each sample analysis costs \$1.

All other things being equal, is the principal investigator better off spending \$9 on laboratory “A” or spending \$9 on laboratory “B”? *Hint – This question is asking you to show me your understanding of standard error.*

**The principal investigator is better off sending his samples to the laboratory that yields the smallest standard error.**

	<b>A</b>	<b>B</b>
<b>Standard deviation, <math>\sigma</math></b>	<b>1 mg</b>	<b>2 mg</b>
<b>Price per analysis</b>	<b>\$3</b>	<b>\$1</b>
<b>\$9 buys # analyses, n =</b>	<b><math>\\$9/\\$3 = 3</math></b>	<b><math>\\$9/\\$1 = 9</math></b>
<b><math>SE = \sigma / \sqrt{n}</math></b>	<b><math>1/\sqrt{3} = 0.577</math></b>	<b><math>2/\sqrt{9} = 0.667</math></b>

**Laboratory A would be preferred, since its data quality includes a smaller standard error**

**9. (10 points total)**

Let “A” and “B” denote two independent genetic traits and suppose that the probability that a randomly selected individual will exhibit trait “A” is  $\frac{1}{2}$ . Suppose also that the probability that a randomly selected individual will exhibit trait “B” is  $\frac{3}{4}$ .

What is the probability that a randomly selected individual will exhibit

- (a) **2 points.** Both traits?

Solution under independence is  $P(A)P(B) = (1/2)(3/4) = 3/8$

- (b) **2 points.** Neither?

By definition, solution is  $1 - P(A \cup B) = 1 - \{ P(A) + P(B) - P(A \text{ and } B) \}$

$$\begin{aligned} & 1 - \{ (1/2) + (3/4) - (3/8) \} \\ & = 1/8 \end{aligned}$$

- (c) **2 points.** Trait A but not trait B?

Solution under independence is  $P(A)P(\text{"not" } B) = (1/2)(1/4) = 1/8$

- (d) **2 points.** Trait B but not trait A?

Solution under independence is  $P(\text{"not" } A)P(B) = (1/2)(3/4) = 3/8$

- (e) **2 points.** Exactly one trait?

$$\begin{aligned} P(\text{"exactly" one trait}) &= P(A \cup B) - P(A \text{ and } B) \\ &= P(A) + P(B) - P(A \text{ and } B) - P(A \text{ and } B) \\ &= 1/2 + 3/4 - 3/8 - 3/8 \\ &= 1/2 \end{aligned}$$

10. (10 points total)

The following are data from a **case-control** study that investigated the relationship between helmet use and facial injury following a bicycle accident.

		Facial Injury		
		Yes	No	
Helmet Worn	No	<b>a = 182</b>	<b>b = 236</b>	418
	Yes	<b>c = 30</b>	<b>d = 83</b>	113
		212	319	531

(a) (2 points) Is this study prospective or retrospective?.

**Retrospective**

(b) (2 points) What are the odds of a facial injury among persons who did not wear a helmet?

**$p=182/418=0.4354$       **Odds= $p/(1-p)=0.7712$****

**Using the letters in the 2x2, odds (injury among non wearers) =  $a/b = 182/236 = .7712$**

(2 points) What are the relative odds (odds ratio) of facial injury for persons who did not wear a helmet compared to those who did wear a helmet?

**Using the letters in the 2x2,**

**odds (injury among non wearers) =  $a/b = 182/236 = .7712$**

**odds (injury among wearers) =  $c/d = 30/83 = .3614$**

**Relative odds (OR) =  $\frac{\text{odds injury among "wearers"}}{\text{odds injury among "nonwearers"}} = \frac{.7712}{.3614} = 2.1339$**

**Using letters in the 2x2 table**

**Relative odds (OR) =  $\frac{\text{odds injury among "wearers"}}{\text{odds injury among "nonwearers"}} = \frac{a/b}{c/d} = \frac{182/236}{30/83} = \frac{(182)(83)}{(30)(236)} = 2.1336$**

- (c) **(2 points)** Is the calculation of the relative risk meaningful in this setting? Explain.

**Strictly speaking, NO. The definition of a case-control study does not permit direct calculation of the relative risk of the disease outcome. This is because subjects are enrolled according to their disease outcome status.**

**However, in the very restricted setting where the disease outcome is a RARE one, a case-control study can be used to obtain a meaningful estimate of the relative risk. This is because, for rare disease outcomes, the OR for the event of history of exposure in a case-control study, which is always identical to the OR for the event of disease in a cohort study, is approximately the same as the RR for the event of disease in a cohort study.**

- (e) **(2 points)** Considering your answers to parts “a” through “d”, in your opinion, does wearing a helmet reduce the risk of sustaining a facial injury in a bicycle accident? Explain.

**Yes. An estimated odds ratio of 2.13 suggests that helmet use is associated with a significantly reduced odds of facial injury compared to the odds of facial injury among persons who do not wear helmets.**