Biostatistics 22S:101 Fall 2005 Practice Final Exam Answers

Name: **ANSWERS** **ID** _____

Posterior Distributions and Statistical Inference.

1. A study comparing coffee drinkers with non-drinkers found that the relative risk (RR) of pancreatic cancer had ***p***-value = .15. Indicate with an "X" which one of the following is consistent with that ***p***-value.

X	95% CI for RR	Significant?
	85 to 3.24	Significant
X	0.85 to 3.24	Not significant
	1.15 to 3.01	Significant
	1.15 to 3.01	Not significant
	-.53 to .67	Not significant

2. Two hundred forty-four male alcoholics suffering from secondary hypertension participated in a study to determine the efficacy of a new antihypertensive agent. The men were assigned at random to either the control group or the treatment group. Men in the control group received a placebo. Statistics for arterial pressure at 30 days after treatment are shown in this table.

Hypertension Study	Placebo	Treatment
n	100	144
mean ($\bar{x}$)	140.3	102
standard deviation (s)	15.0	7.2

Describe the approximate posterior distribution of the difference (Δ) between the means of the two populations and compute a 95% credible interval for the difference.

The posterior distribution is: **Approximately normal with mu = delta-hat = 38.3 and sigma = sed = 1.616**

95% Credible interval: **35.1 to 41.4**

3. In a study of paranoid schizophrenics, the age of onset of each person's illness was determined. Statistics for age at onset are shown in this Table.

Schizophrenia Study	Males	Females
n	9	12
mean ($\bar{x}$)	26.56	29.58
standard deviation (s)	29.78	38.45

Describe the approximate posterior distribution of the difference (Δ) between the means of the two populations and compute a 95% credible interval for the difference.

The posterior distribution is: **approximately t(19) with mu = delta-hat = 3.02 and sigma = sed = 14.89**__

95% Credible interval: **-28.1 to 34.2 (multiplier is 2.09 instead of 1.96)**

Is the difference statistically significant?: No **X** Yes What tells you this?: **CI contains 0** _____

4. A study by Mann et al. investigated whether oral contraceptive use increased the likelihood of heart attacks. The data are in this table,

Contraceptive Study	Used Oral Contraceptives	Never Used Oral Contraceptives
n	57	167
% having heart attacks	23%	35%

Compute the proportions having heart attacks. Describe the posterior distribution of Δ, the difference between the population proportions. Obtain a 95% credible interval for the difference between the population proportions. Is the difference statistically significant? Explain why or why not.

a) Posterior Distribution: **approximately normal with mu = delta-hat = 0.35 - 0.23 = 0.12 and sigma = sed = .067**

b) 95% Credible Interval: **–0.01 to 0.25**

c) Statistically Significant?: Yes ☐ No **X** Why?: **CI includes zero**

5. One possible side effect of antihistamines is drowsiness. In a clinical trial of the drug Suspirizine vs. a placebo . The relative risk of drowsiness in the drug group relative to the placebo group was RR = 8%/6%) = 1.33 and the 95% CI was 1.09 to 1.69. Presumably if a drug produces a "very important" increase in the risk of a side effect the FDA will require a warning label; however we don't know how big the RR has to be to be considered "very important," so let's work two scenarios.

a) The FDA thinks RR > 1.5 is very important. What do you conclude from the study?

You have to reverse-engineer the 95% CI to obtain the posterior distribution of the ln(RR), which is normal with mu = ln(RR-hat) = 0.285 and sigma = ln(UCL/LCL)/(2x1.96) = 0.112. Then you compute the posterior probability that P(RR > 1.5) = P(ln(RR) > .405) = P(Z > (.405 - .285)/.112) = P(Z > 1.08) = 0.14, so there is about a 14% possibility that there is a serious problem.

b) The FDA thinks RR > 2.0 is very important. What do you conclude from the study?

In this case P(RR > 2) = P(ln(RR) > 0.693) = … = 0.000 and the study absolutely rules out the possibility of an important difference.

Statistical Models

6. The patients in the data set in the table below are wearing one of two different cochlear implant devices. Suppose I want a regression model with different intercepts but the same slope for the two devices. In the following spreadsheet fill in columns X1, X2, and X3 with the appropriate design variables so that β_1 is the intercept for device A, β_2 is the intercept for device B, and β_3 is the common slope for the two devices.

Data			Design Variables		
Device	Month 1 (x)	Asymptotic (y)	X1	X2	X3
A	20	31.00	**1**	**0**	**20**
A	35	70.33	**1**	**0**	**35**
A	26	48.75	**1**	**0**	**26**
…	…	…	…	…	…
B	0	25.50	**0**	**1**	**0**
B	22	44.00	**0**	**1**	**22**
B	29	40.00	**0**	**1**	**29**
…	…	…	…	…	…

7. A statistician computed a multiple regression model for the relationship between salary (y), race, gender, and years of experience. This table shows the regression coefficients, their standard errors, p-values, and credible intervals.

				95% Credible Interval	
Variable	$\hat{\beta}$	SE	*p*-value	Lower	Upper
Intercept	30,000	5,000	.0001	20,000	40,000
Male	400	180	[redacted]	40	760
White	700	250	[redacted]	[redacted]	[redacted]
Years	500	60	.001	300	500

What is the average salary for **black males** with 5 years of experience? **32900**

What is the average salary for **white males** with 5 years of experience? **33600**

Find a 95% credible interval for the difference **700 +/- 1.96x250 = 210 to 119**0

Is the difference significant? **Yes** _____

What tells you this? **CI rules out zero** _____

8. The table below shows the results of a (fictional) study of 10-year survival for men who survived a heart attack. The outcome variable is surviving 10 years after first heart attack (1=survived / 0=died), so the logistic regression model predicts the logarithm of the odds on survival. The explanatory variables are family history of early death (1 = at least one first degree male relative died before the age of 55 / 0 = none), body mass index, and age at the time of the first heart attack.

					95% Credible Interval for odds ratio	
	Variable	$\hat{\beta}$	SE	odds ratio	Lower	Upper
1	Intercept	8.1548	0.8864	na	na	na
2	Family History	-0.0513	0.0112	[redacted]		
3	Body Mass Index	-0.1020	0.1627	0.903	0.656	1.242
4	Age	-0.1069	0.0220	1.113	1.066	1.162

Table 1. Logistic regression coefficients for 10-year survival of males following first heart attack.

Compute the probability of survival for a 58-year-old with a family history of early death and BMI = 30 .

log(odds): **–1.157**_____ odds: **exp(–1.157)** = 0.315___ prob: **0.239**_____

Which of the explanatory variables is(are) not statistically significant? What tells you this?

Insignificant variable(s): **BMI**_____

What tells this: **the 95% CI for BMI adjusted odds ratio includes 1.00, the CI for Age rules out 1 and so is significant, the 95% CI for Family Hx has to be computed but the CI for beta for Fam Hx clearly rules out 1.00 which implies that the CI for the adjusted odds ratio (exp(beta)) rules out 1.0 and is therefore significant**

What is the odds ratio for patients with a family history of early death vs. patients without a family history of early death?

Odds ratio: **Adjusted OR = exp(-.0513) = 0.95**_____

Compute a 95% credible interval for β_2 (family history of early death) and for $\exp(\beta_2)$.

Credible interval for β_2: **–.0732 to -0.0293**

Credible interval for $\exp(\beta_2)$: **exp(-.0732) = 0.929 to exp(-0.0293) = 0.971**

9. A study was conducted to determine if maternal smoking influenced the birth weight of infants. Birth weight (pounds) is also influenced by gestational age (number of days between conception and birth). Design variables were the intercept, whether the mother smoked during pregnancy (1=yes/0=no), and the gestational age of the infant (days). The regression results were as follows:

			95% CI	
Variable	$\hat{\beta}$	Std Err	Lower	Upper
Intercept	-1.68	130	[redacted]	
Gestational Age	.052	.022	**0.00888**	**0.09512**
Mother Smoked	-.25	.09	**–0.4264**	**–0.0736**

Compute the credible intervals for the regression coefficients. Does maternal smoking have a significant effect on birth weight? How do you know this?

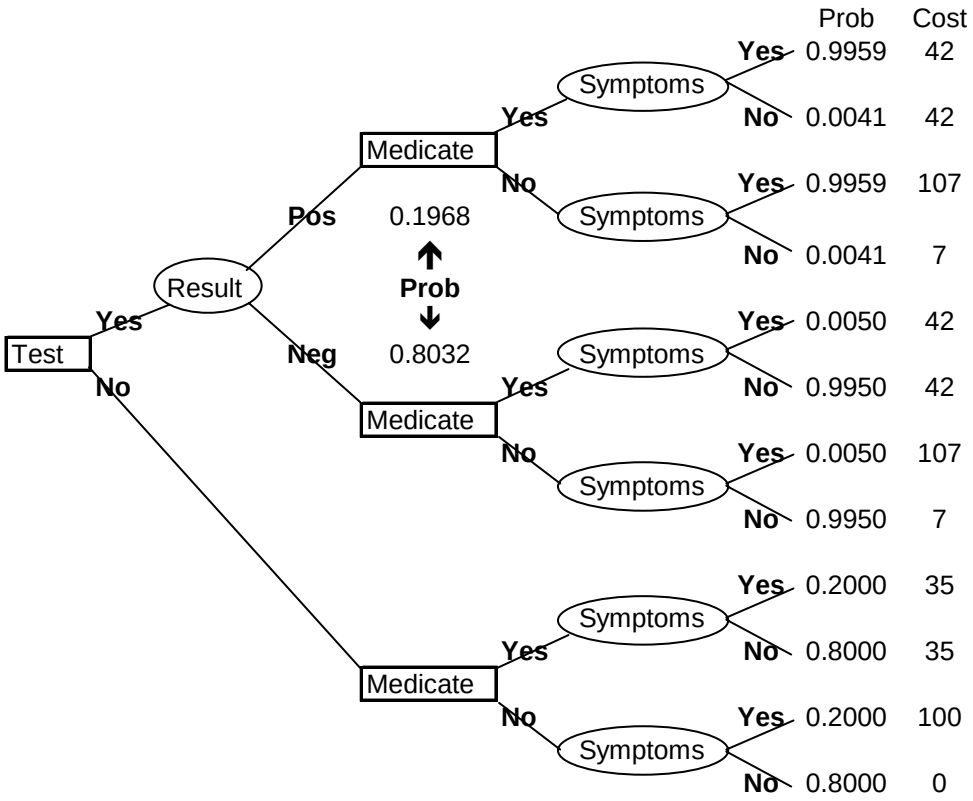
Significant (Y/N)? **Y**_____ Because: **95% CI rules out zero**_____

10. If the credible interval for a logistic regression coefficient (β) rules out zero, what can you say about a) the credible interval for the odds ratio, and b) the p-value?

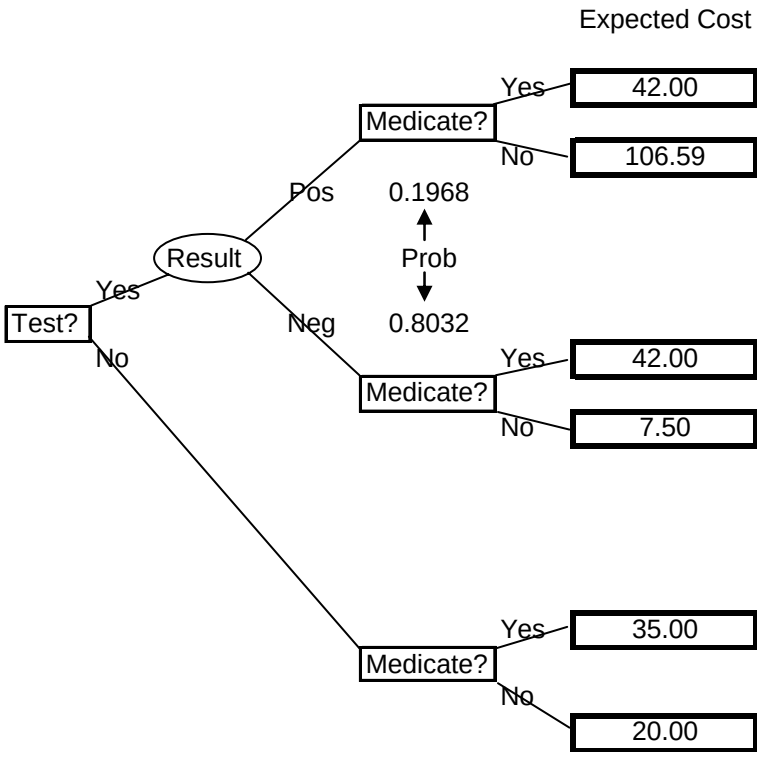
a) The adjusted odds ratio is exp(beta) and the LCL and UCL for adjusted OR are exp(LCL) and exp(UCL). Since LCL > 0 we have exp(LCL) > 1 and since UCL > 0 we have exp(UCL)>1, consequently the CI for the adjusted odds ratio rules out 1.

b) The p-value will be less than 0.05

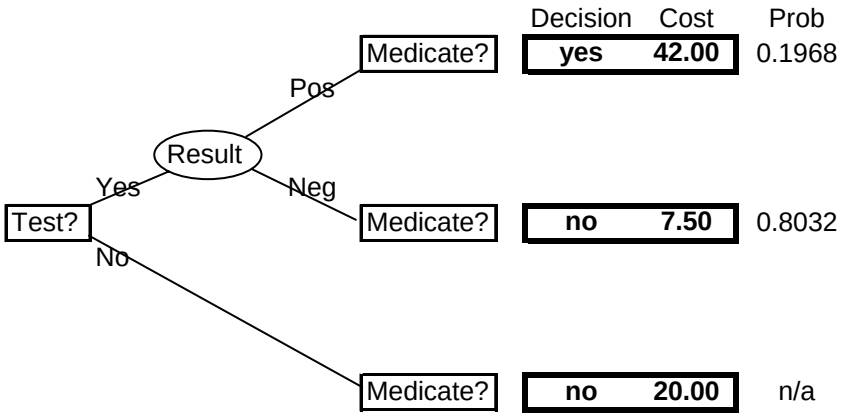
11. The diagram below is a decision problem in tree form. A patient has been exposed to a disease and may or may not be infected but is not yet manifesting symptoms. The physician can call for a diagnostic test or not. The test costs $7 and is not perfect. Based on the test results (or not) the physician can decide to prescribe a $35 prophylactic medication that prevents the disease if it is applied before symptoms appear. If the physician waits for symptoms to appear, it is necessary to prescribe a rescue medication that costs $100. Solve the decision tree.



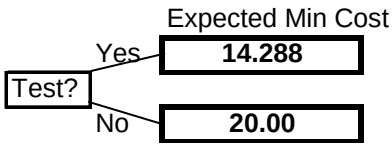
Step 1 Compute Expected costs at the rightmost random nodes:



Step 2, Make the lowest cost decisions at the rightmost decision nodes.



Step 3. Compute the expected cost at the result node.



Step 4. Make the lowest cost decision: TEST

The decision rule: TEST. If the test is positive then medicate, if the test is negative then don't medicate.

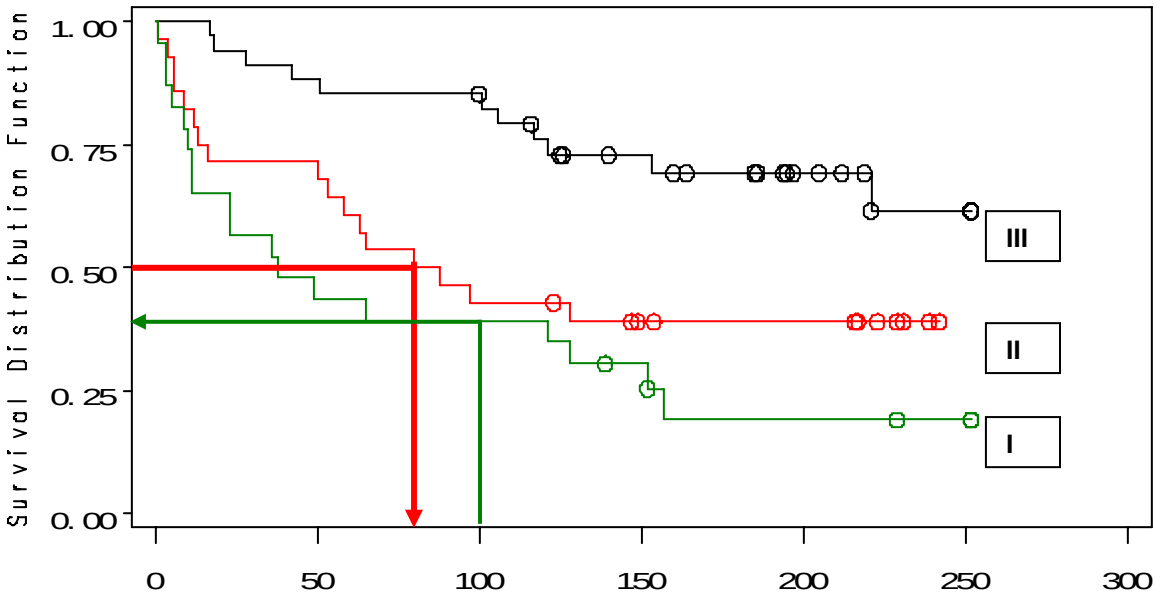
12. The next table shows the calculation of the Kaplan-Meier survival function for these survival times:

1 2 5 6 7 9 15 15 >21 21 >25 >30 31 39 40 >40

Interval	at Risk	Failed	Censored	Failure Rate	Survival Rate
0	n/a	n/a	n/a	n/a	1.0000
0 - 1	16	1	0	0.0625	0.9375
1 - 2	15	1	0	0.0667	0.8750
2 - 5	14	1	0	0.0714	0.8125
5 - 6	13	1	0	0.0769	**0.7500**
6 - 7	12	1	0	0.0833	0.6875
7 - 9	11	1	0	0.0909	0.6250
9 - 15	10	2	0	0.2000	0.5000
15 - 21	8	1	1	**0.125**	0.4375
21 - 25	**6**	0	1	0.0000	0.4375
25 - 30	5	0	1	0.0000	0.4375
30 - 31	4	1	0	0.2500	0.3281
31 - 39	3	1	0	0.3333	0.2188
39 - 40	2	1	1	0.5000	0.1094

Fill in the missing numbers for cells A, B, and C.

13. The following graph shows Kaplan-Meier survival curves for three groups of patients (I, II, and III). Time is expressed in months.



How do the groups compare in terms of survival time, from longest to shortest?.

a) **III>II>I** b) I>II>III c) II>III>I d) II>I>III c) Impossible to determine.

What is the median survival time in Group II? (Draw it on the graph). **Median is about 75 months**

What is the 100-month survival rate for patients in group I? (Draw it on the gr **Survival rate is abut 0.37 or 37%**

14. A sample of n=400 drawn from a fairly normal population has sample mean $\bar{x}$ = 16.5 and sample standard deviation s = 8.8. Describe the approximate posterior distribution of the population mean μ and obtain an approximate 95% confidence interval for μ.

The posterior distribution is: **approximately normal with μ = $\bar{x}$ = 16.5 and σ =sem = 0.44**

95% Credible interval: **16.5 ± 1.96*0.44 or 15.6 to 17.4**

15. An unknown quantity Δ has a t(11) distribution with μ = 20 and σ = 5. Compute the 95% credible interval for Δ. Fill in the blanks in the ttailarea.xls spreadsheet to compute P(Δ > 24). Put a check where you expect the answer to appear.

95% Credible Interval: **20 + 2.20×5 or 9 to 31**

Instructions:
Enter degrees of freedom and z (or t) value.
Read left- or right tail area.

Tail areas for t(df)			
	ν =df =	11	Enter Here
	t=z =	0.80	
Left Tail	P(t(df) ≤ z) =		Read Here
Right Tail	P(t(df) > z) =	✓	