Data Analysis in the Health Sciences

Final Exam 2007 – EPIB–621

Student's Name: ____

Student's Number: ____

INSTRUCTIONS

This examination consists of 8 questions on 16 pages, including this one. Tables of the normal distribution are provided on the last page. Please write your answers (NEATLY) in the spaces provided. Fully explain all of your answers. Each question is worth 10 points, for a total of 80.

1. __________
2. __________
3. __________
4. __________
5. __________
6. __________
7. __________
8. __________

Total (out of 80) __________

1. During the 1980's, many large companies in the United States experimented with "flex-time", in which employees were allowed to choose their own scheduled working hours within limits set by management. Among other things, it was believed that "flex-time" would reduce employee absenteeism. Suppose that the true mean number of days absent per employee per year at a particular company is known to be $\mu_0 = 5.1$. After introducing "flex-time", the number was dropped to $\overline{x} = 4.7$, based on a random sample of 100 employees. The standard deviation for these 100 employees was $s = 3.1$ days.

(a) Test the null hypothesis

H_0: There is no change in absenteeism under "flex-time"

versus the alternative

H_A: There is a change in absenteeism under "flex-time".

Set up the test, calculate a p-value, and state your conclusion.

(b) Calculate a 95% confidence interval for the average number of days absent during “flex-time”. Do you think that the sample size of 100 was adequate? Explain why or why not.

2. In a pilot study measuring the average blood pressure of men aged 50 to 55, the blood pressures of four men in this age group are measured. Their diastolic pressures are found to be 80, 70, 80, and 90 $mm\ Hg$. The investigators of the study state that their prior distribution for the mean blood pressure among such subjects is normally distributed, with mean 75 and variance 25. They also claim to exactly know the standard deviation of blood pressures among men aged 50 to 55 to be 10 $mm\ Hg$.

What is the posterior distribution for the mean blood pressure in this age group, taking into account the prior distribution and the data collected? Provide the mean and variance of this posterior distribution, as well as a rough sketch of this posterior distribution.

3. Data collected over 31 days for the month of January shows that the number of patients arriving at a large emergency room with frostbite per day, Y, is approximately linearly related to the outside temperature on that day, *temp*, measured in degrees Celsius. The minimum temperature during the month was $-40°C$, and the highest temperature recorded during that January was $-5°C$. The estimated equation is:

$$Y = 5 - 0.25 \times temp$$

(a) Give an interpretation for the value of the intercept in the above regression equation.

(b) Give an interpretation for the slope in the above regression equation.

(c) State whether the following statement is true or false, and explain why: "If the temperature is 12 degrees Celsius, the equation predicts that 2 persons on average would come for treatment for frostbite. Since it is almost impossible to get frostbite at such a high temperature, the above equation must contain an error."

4. Alzheimer's Disease is characterized by progressive cognitive impairment, with an average duration of approximately seven years. One way to measure cognitive impairment is through the Mini-Mental State Exam (MMSE). The MMSE ranges from a high of 30 points (no cognitive impairment) to a low of 0 (very severe cognitive impairment). Monthly data were collected on a patient with Alzheimer's Disease who was followed over time. Time was measured in months, with month at diagnosis given a value of 0, and with month 60 representing the end of data collection after five years. A square root transformation was used on the MMSE outcome to achieve linearity.

$$\sqrt{MMSE} = 5 - 0.06 \times month$$

(a) Draw a rough sketch of this regression function, with months on the x-axis, and $MMSE$ (not $\sqrt{MMSE}$) on the y-axis.

(b) Provide a prediction for this subject at month 12, on the original MMSE scale.

(c) Suppose that the 95% confidence interval for the slope of -0.06 is (-0.07, -0.05). State whether the following statement is true or false, and explain why: "A 95% confidence interval for the mean MMSE for this subject at month 12 would be (17.3, 19.4)."

5. A researcher is investigating the effects of independent variables x_1 (a dichotomous variable) and x_2 (a continuous variable) on a dichotomous outcome variable, x_3. The researcher runs the bic.glm program, with the following results:

```
> output<- bic.glm(x3 ~ x1 + x2, glm.family="binomial",
            data=x.data, OR=10000)
> summary(output)

Call:
bic.glm.formula(f = x3 ~ x1 + x2, data = x.data,
          glm.family = "binomial", OR = 1000)


  4  models were selected
 Best  4  models (cumulative posterior probability =  1 ):

       p!=0   EV      SD       model 1     model 2     model 3     model 4
Inter  100   0.06077  0.10108   6.558e-02  -1.279e-02   1.926e-01  -7.605e-03
x1      7.2  0.01490  0.07625       .       1.928e-01       .       4.272e-01
x2     98.9  0.33514  0.09391   3.409e-01   3.094e-01       .           .

nVar                              1           2           0           1
BIC                            -2.423e+03  -2.417e+03  -2.413e+03  -2.412e+03
post prob                       0.922       0.067       0.006       0.005

> output$mle
            [,1]      [,2]      [,3]
[1,]  0.06557595 0.0000000 0.3409430
[2,] -0.01278905 0.1927636 0.3094300
[3,]  0.19259311 0.0000000 0.0000000
[4,] -0.00760460 0.4271544 0.0000000
> output$se
           [,1]      [,2]       [,3]
[1,] 0.09624412 0.0000000 0.08684634
[2,] 0.12459704 0.1949967 0.09247926
[3,] 0.08985735 0.0000000 0.00000000
[4,] 0.12332617 0.1812199 0.00000000
```

(a) Using Bayesian model averaging, what is the optimal prediction for the probability of the outcome when $x_1 = 1$ and $x_2 = 1$?

(b) Discuss any confounding that may be evident from the above results.

(c) Using results from the best (non model averaged) model (i.e., model 1), what is the odds ratio and 95% confidence interval for the odds ratio for the effect of x_2 on the outcome x_3?

6. A group of researchers collects data on height and weight in a random sample of 1000 Canadians, and for each participant, calculates the body mass index (BMI) as $BMI = weight/height^2$. They then create a logistic regression model to predict the probability of low bone mass (a dichotomous variable) using all three variables. Do you think this is a reasonable model? Discuss why or why not.

7. The electrolyte disturbance called hyponatremia exists in humans when sodium concentration in the plasma falls below 135 mmol/L. At lower levels water intoxication may result, an urgently dangerous condition. Hyponatremia has emerged as an important cause of race-related death and life-threatening illness among marathon runners. A cohort of marathon runners was studied to identify the principal risk factors of hyponatremia. Before the race, subjects completed a survey describing demographic information and training history. After the race, runners provided a blood sample and completed a questionnaire. Multivariate logistic regression analyses were performed to identify risk factors associated with hyponatremia. Results were:

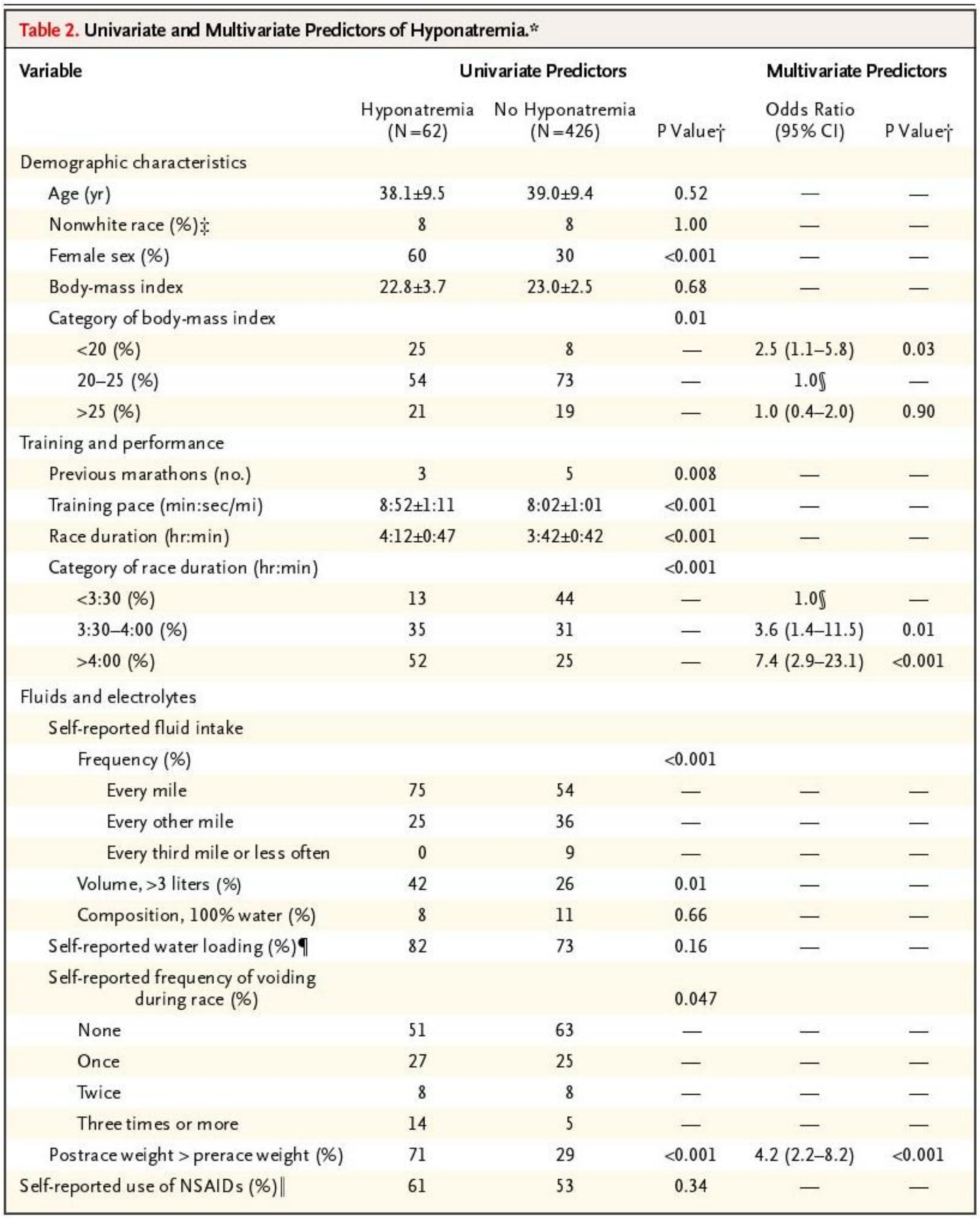

Table 2. **Univariate and Multivariate Predictors of Hyponatremia.***

Variable	Univariate Predictors			Multivariate Predictors	
	Hyponatremia (N=62)	No Hyponatremia (N=426)	P Value†	Odds Ratio (95% CI)	P Value†
Demographic characteristics					
Age (yr)	38.1±9.5	39.0±9.4	0.52	—	—
Nonwhite race (%)‡	8	8	1.00	—	—
Female sex (%)	60	30	<0.001	—	—
Body-mass index	22.8±3.7	23.0±2.5	0.68	—	—
Category of body-mass index			0.01		
<20 (%)	25	8	—	2.5 (1.1–5.8)	0.03
20–25 (%)	54	73	—	1.0§	—
>25 (%)	21	19	—	1.0 (0.4–2.0)	0.90
Training and performance					
Previous marathons (no.)	3	5	0.008	—	—
Training pace (min:sec/mi)	8:52±1:11	8:02±1:01	<0.001	—	—
Race duration (hr:min)	4:12±0:47	3:42±0:42	<0.001	—	—
Category of race duration (hr:min)			<0.001		
<3:30 (%)	13	44	—	1.0§	—
3:30–4:00 (%)	35	31	—	3.6 (1.4–11.5)	0.01
>4:00 (%)	52	25	—	7.4 (2.9–23.1)	<0.001
Fluids and electrolytes					
Self-reported fluid intake					
Frequency (%)			<0.001		
Every mile	75	54	—	—	—
Every other mile	25	36	—	—	—
Every third mile or less often	0	9	—	—	—
Volume, >3 liters (%)	42	26	0.01	—	—
Composition, 100% water (%)	8	11	0.66	—	—
Self-reported water loading (%)¶	82	73	0.16	—	—
Self-reported frequency of voiding during race (%)			0.047		
None	51	63	—	—	—
Once	27	25	—	—	—
Twice	8	8	—	—	—
Three times or more	14	5	—	—	—
Postrace weight > prerace weight (%)	71	29	<0.001	4.2 (2.2–8.2)	<0.001
Self-reported use of NSAIDs (%)‖	61	53	0.34	—	—

The researchers concluded that considerable weight gain while running, a long racing time, and body-mass-index extremes were associated with hyponatremia, whereas female sex, composition of fluids ingested, and use of nonsteroidal antiinflammatory drugs were not associated with hyponatremia.

(a) Based on the table of data given, do you agree with all conclusions as stated above? For each of the six independent variable mentioned in the conclusion, state why you agree or disagree with the conclusions given.

(b) Comparing the univariate continuous results to the univariate categorical results for Body Mass Index, we see that the p-value has changed from 0.68 to 0.01. Explain why this has happened.

8. Suppose that for purposes of the distributions of funds for health services, a province is subdivided into 5 regions. Funds will be distributed in part according to the rates of cancer in each area. A survey is conducted, and data on cancer rates are collected, where x=total number of cancer patients out of n subjects contacted.

Consider the following model, programmed in WinBUGS, and the results which follow:

```
model
{
      for (i in 1:5)
    {

        x[i] ~ dbin(p[i],n[i])
       logit(p[i]) <- z[i]
        z[i] ~ dnorm(mu,tau)
    }

           mu ~  dnorm(0,0.001)
          tau <- 1/(sigma*sigma)
         sigma ~ dunif(0,20)
             y ~ dnorm(mu, tau)
             w <- exp(y)/(1+exp(y))
             p12 <- p[1]-p[2]
}

# Data

list(x=c(60, 50, 30, 40, 50), n=c(1000, 1100, 450, 900, 1100))

# Results

node     mean    sd          MC error    2.5%      median     97.5%  start sample
mu     -2.929   0.1498      0.001519    -3.208  -2.929     -2.65   1001  20000
p[1]    0.0555  0.006405    1.137E-4    0.04494  0.05473    0.0699  1001  20000
p[2]    0.04808 0.005343    7.237E-5    0.03735  0.04828    0.05824 1001  20000
p[3]    0.05699 0.009031    1.529E-4    0.04386  0.05528    0.07862 1001  20000
p[4]    0.04783 0.005698    7.844E-5    0.03612  0.0481     0.05864 1001  20000
p[5]    0.04802 0.00525     7.185E-5    0.03736  0.04815    0.05809 1001  20000
p12     0.00741 0.008296    1.44E-4    -0.00530  0.00595    0.02621 1001  20000
sigma   0.2074  0.219       0.004784    0.00954  0.1551     0.7246  1001  20000
w       0.05314 0.02488     2.107E-4    0.02731  0.05058    0.09219 1001  20000
y      -2.931   0.3416      0.002651    -3.573  -2.932     -2.287   1001  20000
```

(a) Provide a point estimate and 95% credible interval for the rate of a “typical region” drawn from the same distribution of rates as these five regions.

(b) Do you think region 1 has a good case for higher funding compared to region 2? Explain why or why not.

Normal Density Table

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table of standard normal distribution probabilities. Each number in the table provides the probability that a standard normal random variable will be less than the number indicated.