MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II

Friday, August 10, 2018, 9am-3pm.

1. A study investigated potential risk factors for sudden death in women. The risk factors were age (in units of 5 years), history of hypertension (0=no, 1=yes), and history of cigarette smoking (0=no, 1=yes). The study was a case-control study based on a random sample of 500 cases and a random sample of 1000 controls. A multiple logistic regression model was fitted, and partial output is provided below. The response in the model was sudden death (0=control, 1=case). The model fitted was for the logit of the probability that the response is "1".

Variable	Estimate	Standard Error
Intercept	-1.750	0.060
Hypertension (0/1)	0.111	0.031
Smoking (0/1)	0.234	0.072
Age (5 years)	0.100	0.100

   (a) Assess the statistical significance of each individual risk factor.

   (b) What is the odds ratio of sudden death comparing two women who are 10 year apart in age but are similar with respect to the other risk factors. Provide a 95% confidence interval.

   (c) Estimate the probability of sudden death for a 60 year old woman who is a smoker and has a history of hypertension.

   (d) An investigator hypothesized that the effect of smoking on sudden death is the same as the effect of hypertension; the "effect" being the odds ratio. If possible, test this hypothesis using the ouput provided above. If it is not possible, explain why.

   (e) An investigator looked at the distribution of age in the cases and found the highest percentage to be in the age group 50-55 years (age was grouped in 5-year intervals). He concluded that in her lifetime, a woman's risk of sudden death peaks when she is that age (50-55 years). Comment.

   Points: 5 for each part.

2. A study of adolescents with Type 1 diabetes used a behavioral intervention to promote blood sugar control. Participants were randomly assigned to receive the intervention or usual care, and the outcome was whether or not their percent hemoglobin A1c was below 7.5. HbA1c% is a measure of blood sugar control and a value below 7.5 indicates good control (the desirable outcome). Higher values of HbA1c% indicate poor control of diabetes. Of the 150 participants in the study, half were in the intervention group and half were in the control group. After 6 months in the study, 30 of the intervention participants and 19 of the control participants had HbA1c% below 7.5.

   In parts (a)-(c), ignore patient's sex. The data are available in `t1d.dat` and `t2d.dat`.

   Either show your work and calculations or describe your methods and models using common statistical language (not computer code).

   Data summary:

Intervention	Sex	Outcome	COUNT
0	F	0	28
0	F	1	14
0	M	0	28
0	M	1	5
1	F	0	21
1	F	1	20
1	M	0	24
1	M	1	10

   Description of variables:
   Intervention: 1 for intervention, 0 for usual care;
   Outcome: 0 for HbA1c% of 7.5 or higher, 1 for HbA1c% below 7.5;
   Sex: M for Male, F for female;
   Count: Number of patients.

   (a) Compute point estimates and 95% confidence intervals for the probability of good blood sugar control, separately in the control and intervention groups.

   (b) Report a statistic with a 95% confidence interval summarizing the effect of the intervention on blood sugar control. Include one or two sentences appropriately summarizing the results for a non-statistical audience.

   (c) Conduct and report the results of a statistical test to evaluate the effectiveness of the intervention.

   (d) Based on pilot data from similar interventions, the investigator believes the intervention will work better in males than in females. Conduct and report a test examining the

investigator's hypothesis.

(e) Assuming that the intervention has the same effect in males and females, obtain an adjusted (for sex) estimate of that effect and a 95% confidence interval. Include one or two sentences appropriately summarizing the results for a non-statistical audience.

Points: 5 points for each part.

3. The table below gives summary statistics for lung function measurements in random samples from three groups of subjects; non-smokers, light smokers and heavy smoker. Let Y_{ij} denote the lung function measure for the j-th subject in the i-th group. Assume that $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2), i = 1, 2, 3, j = 1, 2, \cdots, n_i$.

i	Group	n_i	Mean	SD
1	Non-smoker	100	3.78	0.79
2	Light smoker	100	3.23	0.86
3	Heavy smoker	100	2.59	0.82

Show your calculations. If you use a computer, simply state that and describe **briefly** what you did using common statistical language. Do **not** give any computer code. Note: The sample standard deviation (SD) above was computed using the formula with $n-1$ in the denominator.

(a) Carry out an ANOVA test of the hypothesis that the lung function means (theoretical population means) in the three groups are all equal. State your hypothesis clearly. Show your calculations and provide the ANOVA table, the test statistic and any associated quantities, and compute the p-value.

(b) Use least-squares to estimate parameters in the model

$$\mu_i = \alpha_1 + \alpha_2 i.$$

Provide the estimates $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2)$. Also provide standard error estimates for $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

(c) Consider the model:

$$\mu_i = \beta_1 + \beta_2 i + \beta_3 i^2.$$

Test the hypothesis $H_0 : (\beta_2, \beta_3) = (0, 0)$ against its complement. Report the test statistic and any associated quantities, and compute the p-value. (Note: H_0 states that both β_2 and β_3 are zero. It is sometimes written as $H_0 : \beta_2 = \beta_3 = 0$).

Points: (a) 9, (b) 8, (c) 8.