Name _____.

Total points: 250

Problems 1-6 (12 points each): Multiple choice questions. Choose a single best answer.

1. For a normal distribution with mean 171 and standard deviation 6, the area between 165 and 177 is about:
   (a) 5% (b) 50% (c) 68% (d) 95%

(c). [For a normal distribution with mean 171 and standard deviation 6, the area between 165 and 177 is from one SD below the mean to one SD above the mean. The area between them is about 68%.]

2. The fraction of girls among newborn babies is about 0.5. Suppose you go to a hospital to record the sex of the newborn babies. Consider the following two events: (i) in the next 100 deliveries, you see 45 or less girls; (ii) in the next 1000 deliveries, you see 450 or less girls. Which is more likely to happen?
   (a) (i) (b) (ii) (c) they are equally likely (d) Need more information to tell

(a). [The observed number of girls follows binomial distributions. For the first event, the distribution of the number of girls has mean 50 with SD = $\sqrt{(100\times0.5\times0.5)} = 5$, and the observed count 45 is just one SD below the mean. For the second event, the distribution of the number of girls has mean 500 with SD = $\sqrt{(1000\times0.5\times0.5)} = 15.8$, and the observed count 450 is >3 SD below the mean. In general, the larger the sample size is, the more centralized the results are.]

3. Suppose the incidence rate of breast cancer is 500 per 100,000 person-years among 40- to 49-year-old premenopausal women with a family history of breast cancer (either a mother or a sister history of breast cancer) (group 1) compared with 200 per 100,000 person-years among 40- to 49-year-old premenopausal women with no family history (group 2). The rate ratio of group 1 versus group 2 is
   (a) 0.005 (b) 0.002 (c) 0.4 (d) 2.5

(d). [The rate ratio is (500/100,000) / (200/100,000) = 2.5.]

4. The level of prostate-specific antigen (PSA) in the blood is frequently used as a screening test for prostate cancer. In a study, the following data were gathered regarding the relationship between a positive PSA test (≥ 4.1 ng/dL) and prostate cancer.

	Prostate cancer		
PSA test results	Yes	No	Total
+	92	27	119
−	46	72	118
Total	138	99	237

   Then the sensitivity of the test is:
(a) 38.8% (b) 50.2% (c) 58.2% (d) 61.0% (e) 66.7% (f) 72.7% (g) 77.3%

(e). [The prostate cancer status is the "truth". The sensitivity of the PSA test is 92/138 = 0.667 = 66.7%.]

5. In a study of birth weight of live-born infants, the birth weight in kilograms (kg) was measured on 200 newborn babies. The 200 observations had average 3.17 and variance 0.198. Then the corresponding coefficient of variation is
   (a) 14.0% (b) 6.2% (c) 712% (d) 16.0%

(a). [$\bar{x} = 3.17$ and the standard deviation is $s = \sqrt{0.198} = 0.445$. Then the coefficient of variation is 0.445/3.17 × 100% = 14.0%.]

6. A study looked at the effects of oral contraceptive (OC) use on heart disease in women 40 to 44 years of age. It found that among 5,000 OC users at the baseline, 13 women developed a myocardial infarction (MI) over a 3-year period, whereas among 10,000 non-OC users, 7 developed an MI over a 3-year period. Then the risk ratio between OC users and non-OC users is about:
   (a) 1.86 (b) 3.71 (c) 0.0019 (d) 0.27

(b). [The risk ratio is (13/5000) / (7/10000) = 3.71.]

Problems 7-15 (12 points each): True/false questions: Select true or false. If false, state the correct result/statement/conclusion that is beyond just grammatically negating a false statement.

7. Standard deviation and inter-quartile range are two different ways of quantifying the spread of a distribution.
   (a) True (b) False

True. [Some of you answered false because of different ways of using the word "spread". In this situation, no points would be deducted.]

8. Correlation coefficient can be used to quantify the level of relatedness between two variables. A correlation coefficient near 1 indicates strong correlation and a correlation coefficient near –1 indicates weak correlation.
   (a) True (b) False

False. A correlation coefficient near 1 indicates strong (positive) correlation; a value near –1 indicates strong (negative) correlation; a value near 0 indicates weak correlation.

9. Logistic regression requires all variables used in data analysis to be categorical, while linear regression requires all variables used in data analysis to be continuous.
   (a) True (b) False

False. Logistic regression only requires the output variable to be binary and linear regression requires the output variable to be continues. Both have no requirements on the types of input variables.

10. The probability of two events happening simultaneously can always be calculated as the product of the two probabilities for the respective events. That is Pr(A and B) = Pr(A) × Pr(B).
    (a) True (b) False

False. This is true only when the two events are independent. In general, Pr(A and B) = Pr(A) × Pr(B given A) or Pr(A and B) = Pr(B) × Pr(A given B).

11. In 1998 the UK government started a campaign to reduce smoking among teenagers from the then national average of 13% to a lower percentage. In 2001, 1000 teenagers were sampled in a survey and 123 were found to smoke. They carried out a z-test to see if the smoking rate in 2001 was different from 13%. The p-value was 0.51. This means the probability of seeing exactly 123 smokers among 1000 teenagers was 0.51.
    (a) True (b) False

False. A p-value is the probability of observing data as extreme as or more extreme than the real data (in our case, ≤ 123 or ≥ 137), not the probability of observing exactly the data, which often is small.

12. In a women's study on the relationship between age of first pregnancy and development of breast cancer, the researchers recorded breast cancer status (yes or no) and age of first pregnancy as five age groups. They carried out a Pearson's chi-squared test to test for independence between these two variables. The degrees of freedom of the test should be five.
    (a) True (b) False

False. The DF is $(2 - 1)\times(5 - 1) = 4$.

13. In a study on the relationship between housing condition and acute lower respiratory infection, the researchers estimated the infection rate ratio of poor housing condition versus good housing condition. The corresponding 95% confidence interval was from 0.59 to 2.32. This means the true rate ratio must be between these two values.
    (a) True (b) False

False. There is a 95% confidence that the interval [0.59, 2,32] contains the true rate ratio. It is possible that the interval doesn't cover the true rate ratio.

14. In a study on the factors influencing the risk of microfilarial infection, the researchers decided to fit a logistic regression with two input variables: area of residence (rainforest or savannah) and age group. In this analysis, the effect of the area of residence on the log-odds of microfilarial infection was assumed to be the same across all age groups and the effect of age group on the log-odds of microfilarial infection was assumed to be the same for both areas of residence.
    (a) True (b) False

True.

15. A hazard function is a function of time. It may have different values at different time points.
    (a) True (b) False

True.

16. (**30 points**) In a study of the level of pulmonary function (measured as forced expiratory volume, or FEV, in liters), researchers fit a linear regression model with four risk factors as input variables: age (in years, ranging from 13 to 19), height (in inches, ranging from 60 to 74), sex (female as 0 and male as 1), and smoking status (non-smoking as 0 and smoking as 1). The Stata output is

```
. regress FEV age height sex smoking

      Source |       SS       df       MS              Number of obs =     117
-------------+------------------------------           F(  4,   112) =   42.64
       Model |  44.3436421     4  11.0859105            Prob > F      =  0.0000
    Residual |  29.1194909   112  .259995455           R-squared     =  0.6036
-------------+------------------------------           Adj R-squared =  0.5895
       Total |   73.463133   116  .633302871           Root MSE      =   .5099

------------------------------------------------------------------------------
         FEV |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .051042   .0289444     1.76   0.081    -.0063075    .1083916
      height |   .0932721   .0191061     4.88   0.000     .0554159    .1311284
         sex |   .6349978   .1318464     4.82   0.000     .3737612    .8962345
     smoking |  -.1328241   .1002904    -1.32   0.188    -.3315367    .0658885
       _cons |  -3.666397   1.262699    -2.90   0.004    -6.168274   -1.164521
------------------------------------------------------------------------------
```

(a) If the model is correct, write down the equation for the outcome, FEV, as a function of age, height, sex, and smoking status.

FEV = –3.666 + 0.051 × age + 0.093 × height + 0.635 × sex – 0.133 × smoking

(b) If the model is correct, what can we say about the effects of the four input variables on FEV and their significance?

FEV increases by 0.051 liter as one grows one year older, if height and smoking status won't change. The effect of age is not significant (it is marginally significant) after adjusting for the effects of height, sex, and smoking.

FEV increases by 0.093 liter as one grows one inch taller, if age and smoking status won't change. The effect of height is significant after adjusting for the effects of age, sex, and smoking status.

FEV is 0.635 liter higher for males than for females with the same age, height, and smoking status. The effect of sex is significant after adjusting for the effects of age, height, and smoking status.

FEV is 0.133 liter lower for smokers than for non-smokers with the same age, height, and sex. The effect of smoking is not significant after adjusting for the effects of age, height, and sex.

(c) If the model is correct, what is the predicted level of FEV for a 17-year-old man who is 65 inches tall and not smoking?

The predicted FEV is –3.666 + 0.051 × 17 + 0.093 × 65 + 0.635 × 1 = 3.881 liter.

17. (**40 points**) A 1985 study identified a group of 509 cancer cases and a group of 489 controls by mail questionnaire. The main purpose of the study was to look at the effect of passive smoking on cancer risk. The study defined passive smoking as exposure to the cigarette smoke of a spouse who smoked at least one cigarette per day for at least 6 months. One potential confounding variable was smoking by the participants themselves (i.e., personal smoking), because personal smoking is related to both cancer risk and spouse smoking. Therefore, it was important to control for personal smoking before looking at the relationship between passive smoking and cancer risk. The data are in the following table:

	Among nonsmokers			Among smokers		
	Case-control status			Case-control status		
Passive smoking	Case	Control	Total	Case	Control	Total
Yes	d_1=120	h_1=80	n_1=200	d_1=161	h_1=130	n_1=291
No	d_0=111	h_0=155	n_0=266	d_0=117	h_0=124	n_0=241
Total	d=231	h=235	n=466	d=278	h=254	n=532

(a) Estimate the odds ratio comparing passive smoking and non-passive smoking after controlling for the effect of personal smoking.

Personal smoking is the stratifying variable. The Mantel-Haenszel method (EMS Chapter 18) can be used to estimate the odds ratio while controlling for the effect of the stratifying variable.

For the first stratum (nonsmokers), $d_1 \times h_0/n = 120 \times 155/466 = 39.91$ and $d_0 \times h_1/n = 111 \times 80/466 = 19.06$. For the second stratum (smokers), $d_1 \times h_0/n = 161 \times 124/532 = 37.53$ and $d_0 \times h_1/n = 117 \times 130/532 = 28.59$. Then $Q = 39.91 + 37.53 = 77.44$ and $R = 19.06 + 28.59 = 47.65$. The Mantel-Haenszel estimate of the odds ratio comparing passive smoking and non-passive smoking is $OR_{MH} = Q/R = 1.63$.

(b) Provide a 95% confidence interval for the odds ratio estimated above.

For the first stratum (nonsmokers), $(d \times h \times n_0 \times n_1)/[n^2 \times (n-1)] = (231 \times 235 \times 266 \times 200)/[466^2 \times 465] = 28.60$. For the second stratum (smokers), $(d \times h \times n_0 \times n_1)/[n^2 \times (n-1)] = (278 \times 254 \times 241 \times 291)/[532^2 \times 531] = 32.95$. Then $V = 28.60 + 32.95 = 61.55$.

The standard error for $\log(OR_{MH})$ is $\sqrt{[V / (Q \times R)]} = \sqrt{[61.55 / (77.44 \times 47.65)]} = 0.129$. The corresponding error factor is $EF = \exp(1.96 \times 0.129) = 1.288$.

Then the 95% confidence interval for the Mantel-Haenszel estimate of the odds ratio comparing passive smoking and non-passive smoking is from $OR_{MH} / EF = 1.63/1.288 = 1.27$ to $OR_{MH} \times EF = 1.63 \times 1.288 = 2.10$.

[Note that some rounding errors may lead to slightly different results in the last digit. Credit won't be deducted for such errors. Big rounding errors would lead to some points deducted.]