**STATISTICS 641 - Exam 1**

**Total time is 100 minutes. The exam is available for 24 hours window starting from noon (CST) February 27, 2014 to 11:59 am (CST) February 28, 2014.**

Name _____

Email Address _____

Please put your answers in the following table.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| | | | | |
| 6 | 7 | 8 | 9 | 10 |
| | | | | |
| 11 | 12 | 13 | 14 | 15 |
| | | | | |
| 16 | 17 | 18 | 19 | 20 |
| | | | | |
| 21 | 22 | 23 | 24 | 25 |
| | | | | |

Please TYPE your name and email address, and **ANSWERS in the above table. This answer table will be graded.** Often we have difficulty in reading the handwritten names and email addresses. Make this cover sheet the first page of your Solutions.

**STATISTICS 641 - Exam #1**

In order to see if Mediterranean diet is superior to a low-fat diet recommended by the American Heart Association the following study was conducted.

The subjects, 605 survivors of a heart attack, were randomly assigned to follow either (1) a diet close to the "prudent diet step 1" of the American Heart Association (control group) or (2) a Mediterranean-type diet consisting of more bread and cereals, more fresh fruit and vegetables, more grains, more fish, fewer delicatessen foods, less meat. For all subjects Canola-oil-based margarine was used instead of butter or cream. The oils recommended for salad and food preparation were canola and olive oils exclusively. Moderate red wine consumption was allowed for all subjects.

Over a four-year period, patients in the experimental condition were initially seen by the dietician, two months later, and then once a year. Compliance with the dietary intervention was checked by a dietary survey and analysis of plasma fatty acids. Patients in the control group were expected to follow the dietary advice given by their physician. The researchers collected information on number of deaths from cardiovascular causes e.g., heart attack, strokes, as well as number of nonfatal heart-related episodes. The occurrence of malignant and nonmalignant tumors was also carefully monitored. Answer the next two questions.

1. This study is an example of a

   (a) stratified cluster random sample survey.

   (b) stratified random sample survey.

   (c) retrospective study.

   (d) prospective study. •

   (e) cross-sectional study.

   Explanation: Subjects were followed over time. During or in the end of the follow-up the researchers gathered information if someone has a cardiovascular disease.

2. The outcome and the potential predictors are

   (a) heart attack and canola oil.

   (b) cardiovascular disease and two diet groups, experimental and control. •

   (c) Red wine and two diet groups, experimental and control.

   (d) Oil and two diet groups, experimental and control.

3. In order to study the academic performance of high school students in the United States, a researcher chose a random sample of 15 states and then from each selected states he randomly selects 5 high schools, and from each selected school he then randomly samples 100 students. This is an example of

   (a) a simple random sample.

   (b) a stratified random sample.

   (c) a cluster random sample.

   (d) a stratified cluster random sample.

   (e) a multistage cluster random sample. •

   Explanation: Among the 50 states, 15 states are randomly drawn. Each state is a cluster of many high schools. From each selected states 5 high schools were randomly selected. Each high school is a cluster of students. From each selected high schools 100 students were randomly selected. Therefore, it is a multistage cluster sampling.

   A nuclear engineer has determined that failures in the back up pumps in a nuclear power plant occur according to a Poisson process with an average rate of 2 failures per month. She wants to simulate the lengths of times, $T$, between consecutive emergencies. Answer the next two questions.

4. What is $\text{pr}(T > 4)$?

   (a) 0.0015

   (b) 0.00004

   (c) 0.00865

(d) 0.65

(e) 0.0003 •

Explanation: $T \sim \text{Exponential}(\mu = 0.5)$, $\text{pr}(T > 4) = \exp\{-4(2)\} = \exp(-8)$.

5. What is the probability that there is no pump failure during a month?

   (a) 0.254

   (b) 0.152

   (c) 0.271

   (d) 0.864

   (e) 0.135 •

   Explanation: $\text{pr}(X = 0) = \exp(-2)$

   A car insurance company wants to model the number of claims $X$ filed by each customer. Answer the next two questions.

6. Suppose that it is also known that mean and variance of $X$ are quite different. A reasonable model for $X$ is

   (a) Binomial.

   (b) Hypergeometric.

   (c) Negative Binomial. •

   (d) Poisson.

   (e) Beta.

   Explanation: Number of claims $X$ is a count variable. Since the mean and variance of $X$ are quite different, a Poisson model would not be appropriate. Among the given choices Negative Binomial distribution is the best choice.

7. Suppose that $X$ follows a Poisson distribution with parameter 0.55. What is the probability that there is at least 3 claims in total filed by 10 customers?

   (a) 0.236

   (b) 0.764

   (c) 0.472

   (d) 0.324

   (e) 0.912 •

   Explanation: $Y$ = total claim by 10 customers, $Y \sim \text{Poisson}(10 \times 0.55 = 5.5)$, $\text{pr}(Y \geq 3) = 1 - \text{pr}(Y \leq 2)$

8. Skewness of a distribution is measured using

   (a) the second and third moments. •

   (b) the first and third moments.

   (c) the third and fourth moments.

   (d) the second and fourth moments.

   (e) the second and fifth moments.

9. The lifetime of a fluorescent bulb follows an Exponential distribution with mean 5 years. The total lifetime of 5 such light bulb will follow

   (a) an Exponential distribution.

   (b) a Uniform distribution.

   (c) a Gamma distribution. •

   (d) a Normal distribution.

(e) a Poisson distribution.

Explanation: Sum of independent and identically distributed Exponential random variables follows a Gamma distribution

10. Suppose that $X \sim \text{Binomial}(5, 0.4)$. Find the 40th quantile of the distribution.

    (a) 0
    (b) 1
    (c) 2 •
    (d) 3
    (e) 4

    Explanation: $F(0) = 0.07776$, $F(1) = 0.33696$, $F(2) = 0.68256$, $F(3) = 0.91296$. Thus $Q(0.40) = \inf\{x : F(x) \geq 0.40\} = \inf\{2, 3, 4, 5\} = 2$

11. Consider an AR(1) process $X_t = \theta + \rho X_{t-1} + \epsilon_t$, $\epsilon_t \sim \text{Normal}(0, \sigma_e^2)$, with $\rho > 0$. Get the expression for the variance of $X_t$.

    (a) $\sigma_e^2/(1-\rho)$
    (b) $\sigma_e^2/(1+\rho)$
    (c) $\sigma_e^2/(1+\rho^2)$
    (d) $\sigma_e^2/(1-\rho^2)$ •

    Explanation: HW problem.

    Suppose that in an observational study we observe $V_i, \Delta_i, W_i$, $i = 1, \cdots, n$ where $V_i$ denotes the minimum of the actual age of onset ($T$) of heart disease and a random censoring time ($C$) with $T_i$ and $C_i$ independent. The censoring indicator is denoted by $\Delta_i$. That means $V_i = \min(T_i, C_i)$ and $\Delta_i = 1$ if $T_i \leq C_i$ and 0 otherwise. Here $W_i$ denotes the weight of the $i^{th}$ subject. Answer the next two questions.

12. Assume that $T$ follows the Exponential distribution with mean $\lambda$ and $C$ follows the Exponential distribution with mean 1. Based only on the data $(V_1, \cdots, V_n)$ the method of moment estimator of $\lambda$ is (Hint: The distribution of $V_i$ is exponential with mean $\lambda/(1+\lambda)$)

    (a) $\overline{V}$
    (b) $1/(\overline{V} - 1)$
    (c) $1/\overline{V}$
    (d) $\overline{V}/(1-\overline{V})$ •
    (e) Cannot be determined.

    Explanation: In the method of moment method we equate the fist population moment with the first sample moment, if there is only one parameter. Here we equate first sample moment $\overline{V}$ with $E(V) = \lambda/(1+\lambda)$, i.e., $\overline{V} = \lambda/(1+\lambda)$, and solve for $\lambda$.

13. Suppose the study contains 30% female and 70% male. The distribution of weight of the subjects is most likely to be a

    (a) mixture of two Poisson distributions
    (b) Binomial distribution
    (c) Gamma distribution
    (d) mixture of two Beta distributions
    (e) mixture of two Normal distributions •

14. Suppose that $U = 0.52$ is a random number from a Uniform distribution. Use this random number to generate a random number from the Exponential distribution with mean 2.

    (a) 1.307853

(b) 2.056556

(c) 0.326963

(d) 0.366984

(e) 1.467938 •

Explanation: For exponential distribution with mean 2, the CDF is $F(x) = 1 - \exp(-x/2)$. Now we set $0.52 = F(x) = 1 - \exp(-x/2)$ and solve for $x$. We get $\exp(-x/2) = 0.48$, so $x = -2\log(0.48)$.

15. Suppose that the national quarterly unemployment rate follows a first order moving average model

$$Y_t = \beta\epsilon_{t-1} + \epsilon_t,$$

where $\epsilon_t$ are iid, $E(\epsilon_t) = 0$, $\text{var}(\epsilon_t) = 2.8$, and $\epsilon_t$ are independent. Find the correlation$(Y_t, Y_{t-1})$.

(a) $\beta$

(b) $\beta/(1+\beta^2)$ •

(c) $(1+\beta^2)$

(d) $\beta/(1+\beta)$

(e) $(1-\beta)/(1+\beta)$

Explanation: $\text{var}(Y_t) = \beta^2(2.8) + (2.8)$ for all $t$. $\text{cov}(Y_t, Y_{t-1}) = \text{cov}(\beta\epsilon_{t-1} + \epsilon_t, \beta\epsilon_{t-2} + \epsilon_{t-1}) = \beta\text{cov}(\epsilon_{t-1}, \epsilon_{t-1}) = \beta\text{var}(\epsilon_{t-1}) = \beta(2.8)$. Thus, correlation$(Y_t, Y_{t-1}) = \text{cov}(Y_t, Y_{t-1})/\sqrt{\text{var}(Y_t)\text{var}(Y_{t-1})} = \beta/(1+\beta^2)$.

16. The purpose of autocorrelation plot is

(a) to compute the correlation between a response and explanatory variable.

(b) to check if the model errors and the explanatory variables are correlated.

(c) to check if the response variable is dependent on the its values at the previous periods. •

(d) to check if the errors are correlated.

(e) none of the above.

17. In the development of a new treatment for kidney disease in domestic cats, 100 cats with kidney problems are placed on a new treatment. The time $T$ until the cat no longer has kidney disease is recorded for each of the 100 cats. A plot of the hazard function yields $h(t) = 2t^2$. Find $\text{pr}(T \geq 1)$. [Hint: The survival function satisfies $S(t) = \exp\{-\int_0^t h(u)du\}$.]

(a) 0.5134 •

(b) 0.4174

(c) 0.5825

(d) 0.0799

(e) 0.4765

Explanation: $S(1) = \text{pr}(T \geq 1) = \exp(-2\int_0^t u^2 du) = \exp(-2/3) = 0.5134$

Let two random variables $U_1$ and $U_2$ both follow Uniform$(0, 1)$ distribution. Assume that $U_1$ and $U_2$ are independent. Answer the next two questions.

18. Find the probability that $U_1 + U_2 < 0.5$.

(a) 0.125 •

(b) 0.25

(c) 0.30

(d) 0.75

(e) 0.625

Explanation: Draw a square so that each side has length 1. Now draw the straight $y + x = 0.5$. The area within the square and this line is the probability of the event $U_1 + U_2 < 0.5$. This area is right angle triangle with with base and height equal to 0.5. Thus the area of the triangle is $0.5(0.5)/2 = 0.125$.

19. What is the distribution of $U_1^2 + U_2^2$?

    (a) Exponential distribution

    (b) Normal distribution

    (c) Chi square distribution

    (d) Uniform distribution

    (e) None of the above •

    Explanation: Let $Y = U_1^2 + U_1^2$. Cannot be a) because $Y$ has a bounded support $[0, 1]$ whereas exponential distribution has support $[0, \infty)$. Cannot be b) because normal distribution has support $(-\infty, \infty)$. Cannot be c) because Chi-square distribution has support $[0, \infty)$. Observe that for $0 < x <\leq 1$, $\text{pr}(U_1^2 \leq x) = \text{pr}(U_1 \leq \sqrt{x}) = \sqrt{x} > x = \text{pr}(U_1 \leq x)$. Thus the distribution of $U_1^2$ is no longer a uniform, rather it is a right skewed distribution. The same is true for $U_2^2$. Thus, there $Y$ cannot follow a uniform distribution.Hence d) cannot be the correct answer.

20. A relative frequency histogram having classes of greatly different class widths was used as an estimator of a continuous population pdf. The relative frequency was plotted versus the class intervals. The plot will result in a graphical distortion. The plot can be corrected by

    (a) making all the intervals have the same width.

    (b) increasing the sample size.

    (c) making sure that the area under the curve adds to one

    (d) plotting the relative frequency divided by class width. •

    (e) In fact there will not be a distortion since it is an unbiased estimator of the pdf.

    Suppose that we have a hypothetical set of observations

    $$2, 15, 18, 20, 21, 40, 42, 56, 65, 90, 91, 100, 110+,$$

    from an unknown distribution $F$, where + denotes a right censored observation. Answer the next two questions.

21. Obtain the nonparametric estimate of $\text{pr}_F(X > 90)$.

    (a) 3/13 •

    (b) 2/12

    (c) 3/12

    (d) 4/13

    (e) 9/13

    Explanation: # subjects $> 90$/total number of subjects $= 3/13$.

22. Suppose that the estimate of $\text{pr}_F(X > 90)$ is 0.20 (this is not actually true, but take it for the time being). What is the estimated standard error of your estimator?

    (a) 0.049

    (b) 0.057

    (c) 0.115

    (d) 0.111 •

    Explanation: $\sqrt{(}0.2 \times 0.8/13)$

23. An experiment involves putting specimens of steel under stress until the specimen fractures. The machine increases the stress until the specimen fractures. The maximum stress that the machine can place on a specimen is 500 psi. Out of the 35 specimens used in the experiment, 5 did not fracture at 500 psi. This type of censoring is called

(a) Random censoring

(b) Type I censoring •

(c) Type II censoring

(d) Left censoring

Answer the next two questions based on the following CDF of a hypothetical discrete random variable $X$.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.2 & \text{if } x = 0 \\ 0.48 & \text{if } x = 1 \\ 0.65 & \text{if } x = 2 \\ 0.84 & \text{if } x = 3 \\ 0.99 & \text{if } x = 4 \\ 1 & \text{if } x \geq 5. \end{cases}$$

24. Find the IQR of $X$.

(a) 4

(b) 3

(c) 2 •

(d) 1

(e) 0

Explanation: $Q(0.75) = 3, Q(0.25) = 1$, so $IQR = 2$

25. Let $X_1$ and $X_2$ be two independent observations with both having distribution $F$. What is the probability that $X_1 + X_2 \leq 1$?

(a) 0.192

(b) 0.096

(c) 0.220

(d) 0.462

(e) 0.232 •

Explanation: $\text{pr}(X_1 = X_2 = 0) + \text{pr}(X_1 = 1, X_2 = 0) + \text{pr}(X_1 = 0, X_2 = 1) = 0.2 \times 0.2 + 2 \times 0.2 \times 0.48$

26. **Bonus question:** Consider two distributions, i) Normal$(\theta_1, 1)$ and ii) Lognormal$(\theta_1, \theta_2)$. Which of the following statements is correct?

(a) i) and ii) both belong to location family

(b) i) belongs to a location family but not ii) •

(c) ii) belongs to a location family but not i)

(d) Neither i) nor ii) belongs to a location family