**PH 538: Biostatistical Methods I**

**HOMEWORK 2 (Probability and Sampling distribution) [Due October 6, 2016]**

**Dr. Fares Qeadan: fqeadan@salud.unm.edu**

|  |
| --- |
|  |

**Question 1[1 points]:**

**Which of the following statements is correct?**

(a) The term probability refers to the study of certainty.

(b) *A* and *B* are said to be independent events if *P (A|B) =P (B|A)*.

(c) When events *A* and *B* have no outcomes in common, they are said to be mutually exclusive or disjoint events.

(d) None of the above.

**Answer:**

**Question 2[1 points]:**

**We use the notation ……………….. to represent the conditional probability of event** *A* **given that event** *B* **has occurred.**

(a) *P* (*B*|*A*)

(b) *P* (*A*|*B*)

(c) *P* (*A*Ո*B*)

(d) *P* (*A*Ս*B*)

(e) *P* (*A*|*Bc*)

**Answer:**

**Question 3[1 points]:**

**Two events** *A* **and** *B* **are ……………if** *P* (*A*|*B*) = *P* (*A*)**.**

(a) Disjoint

(b) Complementary

(c) Mutually exclusive

(d) independent

**Answer:**

**Question 4[1 points]:**

**If** *P* (*A*) = 0*.*35 **and** *P* (*B*) = 0*.*45**, then** *P* (*A*Ո*B*)

(a) is .10

(b) is .80

(c) is .20

(d) Cannot be determined from the given information.

**Answer:**

**Question 5[5 points]:**

For the events A and B

1. If P (A) = .50, P (B) = .70, and P (A Ո B) = .40, then P (A Ս B) =............
2. If P (A) = .70, P (B) = .60, and P (A Ս B) = .90, then P (A Ո B) =............
3. If P (A) = .35, P (B) = .50 and P (A Ո B) = .20, then P (A|B) =............
4. If P (A) = .60, P (B) = .30, and P (A|B) = .50, then P (A Ո B) =............
5. If P (A Ո B) = .123 and P (A|B) = .60, then P (B) =............

**Question 6[6 points]:**

A survey of patients at a hospital finds that 40% like to be served low-sodium food, 30% like to be served fried food, and 10% like to be served both.

1. What percent of patients like to be served either low-sodium or fried food?

**Answer:**

1. What percent of patients like to be served low-sodium but not fried food?

**Answer:**

1. What percent like to be served neither low-sodium nor fried food?

**Answer:**

1. What is the chance that a randomly selected patient likes low-sodium food given that he/she likes fried food?

**Answer:**

1. What is the probability that a randomly selected patient likes to be served only one of the two types of food?

**Answer:**

1. Are liking to be served low-sodium food and liking to be served fried food independent events? Why or why not?

**Answer:**

**Question 7[2 points]:**

The overall prevalence of H. pylori infection in Australia is about 20%, with no statistical difference between genders. In fact, almost everywhere, H. pylori infection has already been shown to be independent of gender. Assume that the proportion of males in Australia is 50%, then the probability that a randomly selected individual from Australia will either be a male or an individual having H. pylori is 60%. Explain why by showing your work.

**Answer:**

**Question 8[5 points]:**

Consider the following natality statistics for NM population in 2010. According to these data, the probabilities that a randomly selected women, from NM, who gave birth in 2010 was in each of the following age groups are as follows:

|  |  |
| --- | --- |
| Age | Probability |
| Under 15 | 0.00166 |
| 15-19 | 0.13906 |
| 20-24 | 0.29449 |
| 25-29 | 0.27236 |
| 30-34 | 0.18306 |
| 35-39 | 0.08725 |
| 40-44 | 0.02051 |
| 45-49 | 0.00140 |
| 50-54 | 0.00022 |
| Total | 1.00000 |

1. What is the probability that a randomly selected women, from NM, who gave birth in 2010 was at least 25 years of age?

**Answer:**

1. What is the probability that a randomly selected women, from NM, who gave birth in 2010 was at most 19 years of age?

**Answer:**

1. Given that the mother of a particular child was under 30 years old, what is the chance that she was not yet 20?

**Answer:**

1. Given that the mother of a particular child was 35 years of age or older, what is the chance that she was under 40?

**Answer:**

1. What is the probability that a randomly selected women, from NM, who gave birth in 2010 was between 30 and 39 years of age inclusive?

**Answer:**

**Question 9[11 points]:**

Consider the following natality statistics for NM population in 2010 by race. The provided table shows the Number of Live Births by Mother's Race and Ethnicity and Mother's Age Group:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Mother's Race and Ethnicity** | | | | | | **Total** |
| **American Indian or Alaska Native** | **Asian or Pacific Islander** | **Black** | **Hispanic** | **White** | **Others** |
| **Mother's Age Group** | **Under 15** | 9 | 0 | 0 | 32 | 5 | 0 | 46 |
| **15-19** | 523 | 15 | 67 | 2600 | 642 | 18 | 3865 |
| **20-24** | 1,086 | 61 | 176 | 4,827 | 1,986 | 49 | 8185 |
| **25-29** | 858 | 170 | 112 | 3,985 | 2,407 | 38 | 7570 |
| **30-34** | 555 | 168 | 92 | 2,360 | 1,889 | 24 | 5088 |
| **35-39** | 319 | 113 | 31 | 1,092 | 856 | 14 | 2425 |
| **40-44** | 62 | 19 | 8 | 264 | 212 | 5 | 570 |
| **45-49** | 3 | 0 | 2 | 13 | 21 | 0 | 39 |
| **50-54** | 0 | 0 | 0 | 0 | 6 | 0 | 6 |
| **Total** | | 3415 | 546 | 488 | 15173 | 8024 | 148 | 27794 |

1. The probability distribution for women in NM during 2010 by maternal age was provided in problem 8. Could one extracts the very same probability distribution from the above table? What such probability distribution is called?

**Answer:**

1. Construct the marginal probability distribution for Mother’s Race and Ethnicity?

**Answer:**

1. What is the probability that a randomly selected women, from NM, who gave birth in 2010 was Hispanic (i.e. what is the empirical probability of Hispanic mothers)?

**Answer:**

1. What is the conditional probability of Hispanic mothers given that Mother’s maternal age is between 15-19 years old?

**Answer:**

1. Are maternal age of 15-19 and Hispanic race independent events? Why or why not? (round off your answers to 3 decimal points)

**Answer:**

1. What is the probability that a randomly selected women, from NM, who gave birth in 2010 was from other races (i.e. what is the empirical probability of Others)?

**Answer:**

1. What is the conditional probability of Others given that Mother’s maternal age is between 25-29 years old?

**Answer:**

1. Are maternal age of 25-29 and Others race independent events? Why or why not? (round off your answers to 3 decimal points)

**Answer:**

1. Construct the conditional probability distribution for the maternal age group 25-29?

**Answer:**

1. Given that the mother’s maternal age was 50-54, what is the chance that the mother’s race and ethnicity was not White?

**Answer:**

1. Given that the mother’s race and ethnicity was not White, what is the chance that the mother’s maternal age was 50-54?

**Answer:**

**Question 10[1 points]:**

Which one of the following is implied by a screening test with a low sensitivity?

a. The test yields a large number of false positives.

b. The test yields a large number of false negatives.

c. The test yields a large number of true positives.

d. The test yields a large number of true negatives.

**Answer:**

**Question 11[1 points]:**

When the area under the ROC curve is 0.5

a. The diagnostic test is very helpful.

b. The diagnostic test is somewhat helpful.

c. The diagnostic test is not helpful.

d. It cannot be determined from the information given.

**Answer:**

**Question 12[1 points]:**

Which of the following characterizes a good screening test?

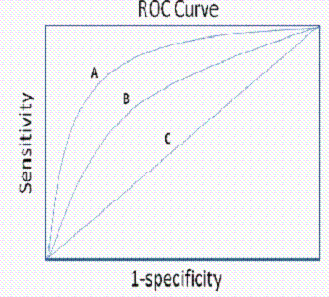
a. high sensitivity and high specificity c. low sensitivity and high specificity

b. high sensitivity and low specificity d. low sensitivity and low specificity

**Answer:**

**Question 13[2 points]:**

Which (A, B, or C) represents the best screening test? Why? What does curve C tell about the screening test?



**Answer:**

**Question 14[2 points]:**

In a body mass index (BMI) test, a cutoff point is established such that an individual is classified as obese when test score is above the cutoff point. As the cutoff score increases, how will the sensitivity and specificity of the test change? (You may want to read the following document about the effect of cut-off point on sensitivity and specificity before answering the question: <http://www.mathalpha.com/PH-538/cutoff_point.pdf> )

**Answer:**

**Question 15[4 points]:**

In NM, 9.4% of the population are Natives, 46.3% are Hispanics, 40.5% are Whites, 2.5% are Black/African American, and 1.3% are Asian/Pacific Islanders. The rate of hypertension by race is known to be as follows: 24.0% of Natives have hypertension, 26.7% of Hispanics have hypertension, 33.1% of Whites have hypertension, 40.8% of Black/African American have hypertension, and 21.3% of Asian/Pacific Islander have hypertension.

1. What is the rate of Hypertension in NM?

**Answer:**

1. Given that an individual from NM is having hypertension, what is the chance that this individual is Native?

**Answer:**

**Question 16[11 points]:**

A screening test for a newly discovered disease is being evaluated for its effectiveness and sensitivity as a screening test in industry. In order to determine the effectiveness of the new test, it was administered to 800 subjects. Of those, 100 of the individuals diagnosed with the disease tested positive for it. Results from the test showed a negative test finding for 50 people with the disease. A total of 55 people without the disease tested positive for it.

1. Create a 2x2 table for this data:

|  |  |  |  |
| --- | --- | --- | --- |
|  | D+ | D- | Total |
| T+ |  |  |  |
| T- |  |  |  |
| Total |  |  |  |

1. What is the sample prevalence rate of the disease?

**Answer:**

1. What is the sensitivity of the test?

**Answer:**

1. What is the specificity of the test?

**Answer:**

1. What is the percentage of false negatives?

**Answer:**

1. What is the percentage of false positives?

**Answer:**

1. What is the accuracy of the test?

**Answer:**

1. What is the PPV of the test?

**Answer:**

1. What is the NPV of the test?

**Answer:**

1. What is the likelihood ratio of the test?

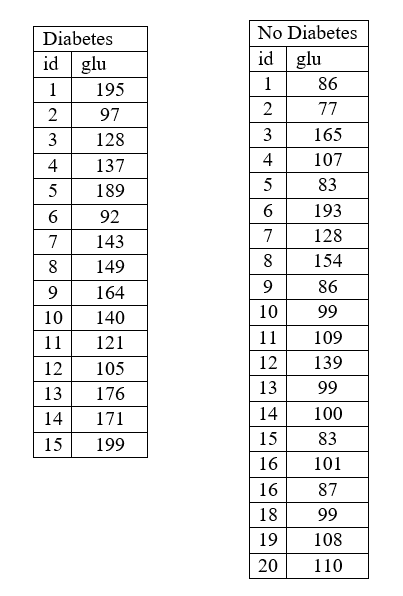
**Answer:**

1. Is this a good test? Use your previous answers from (2) to (10) to justify your answer.

**Answer:**

**Question 17[8 points]:**

There are numerous tests used for diabetes diagnosis including fasting plasma glucose (FPG), random plasma glucose (RPG), oral glucose tolerance test (OGTT) and the A1C test. In the FPG test, a cut-off of 126 mg/dL confirms a diagnosis of diabetes such that anyone with glucose level (glu) equal to or larger than 126 mg/dL is diagnosed with diabetes, but with glucose level less than 126 mg/dL is not diagnosed with diabetes. Information, from an FPG test, on glucose level for a sample of 15 diabetic and 20 non-diabetic patients is shown below.



1. Using the data above, what are the estimated sensitivity and specificity of the FPG diagnostic test?

**Answer:**

1. Prevalence of diabetes is about 10%. Using the estimated sensitivity and specificity in part (a), calculate the estimated probability that a randomly selected subject does not actually have diabetes given that he/she was diagnosed with diabetes?

**Answer:**

1. What is the sample prevalence of diabetes? Compute the positive predictive value firstly from the sample data and secondly by assuming a population prevalence of 10%? Which answer is more accurate? Explain

**Answer:**

1. How would the estimated sensitivity and specificity change if the cut-off point was increased from 126 to 130 mg/dL? Explain (you may want to read the following document about the effect of cut-off point on sensitivity and specificity before answering the question: <http://www.mathalpha.com/PH-538/cutoff_point.pdf> )

**Answer:**

1. Compute the likelihood ratio and interpret it?

**Answer:**

1. What is the probability of this test failing to detect diabetes?

**Answer:**

1. What is the probability of this test detecting diabetes when it’s not present?

**Answer:**

1. How would you interpret the results of this diagnostic test if you tested positive? If you tested negative?

**Answer:**

**Question 18[1 points]:**

Suppose that the serum levels of total cholesterol y for a sample of adult males is bell-shaped with mean 180 and standard deviation 10. The distribution is sketched below.



1. Approximately what percentage of sample males have y-values between 160 and 210? (Please use the 68-95-99.7% Rule to justify your answer and not STATA)

**Answer:**

1. Mr. Smith is told that his cholesterol level y is located 2.5 standard deviations below the mean, that is, has a z-score = -2.5. What is his cholesterol level?

**Answer:**

**Question 19[1 points]:**

The Central Limit Theorem says the sampling distribution of the sample mean is approximately normal under certain conditions. What is a necessary condition for the Central Limit Theorem to be used?

1. The population size must be large (e.g., at least 30).
2. The population from which we are sampling must not be normally distributed.
3. The population from which we are sampling must be normally distributed.
4. The sample size must be large (e.g., at least 30).

**Answer:**

**Question 20[1 points]:**

Which of the following statements about the sampling distribution of the sample mean is incorrect?

1. The standard deviation of the sampling distribution is σ.
2. The mean of the sampling distribution is μ.
3. The sampling distribution is approximately normal whenever the sample size is sufficiently large (n ≥ 30).
4. The sampling distribution is generated by repeatedly taking samples of size n and computing the sample means.

**Answer:**

**Question 21[0.5 points]:**

Any random variable whose only possible values are 0 and 1 is called a ……………… random variable; a special name given after the individual who first studied it.

(a) Bernoulli (b) Poisson (c) Gaussian (d) Laplace (e) Cauchy

**Answer:**

**Question 22[1 points]:**

Which of the following is (are) conditions of a binomial experiment?

1. There is a sequence of n trials, where n is fixed in advance of the experiment.
2. The trials are identical, and each trial can result in one of the same two possible outcomes, which we denote by success (S) or failure (F).
3. The trials are independent, so that the outcome of any particular trial does not influence the outcome of any other trial.
4. The probability of success is the same (constant) from trial to trial; we denote this probability by p.
5. All of the above are conditions of a binomial experiment.

**Answer:**

**Question 23[0.5 points]:**

If X is a continuous random variable, and c is any number, then P(X = c) =…………….

(a) 0.00 (b) 0.25 (c) 0.50 (d) 0.75 (e) 1

**Answer:**

**Question 24[8 points]:**

About 5% of pregnant women develop a form of type 2 diabetes, usually temporary, in their third trimester called gestational diabetes. If 20 pregnant women are selected at random, using STATA:

1. What is the probability that exactly 3 of them will develop the disease?

**Answer:**

1. What is the probability that less than 3 of them will develop the disease?

**Answer:**

1. What is the probability that at most 3 of them will develop the disease?

**Answer:**

1. What is the probability that at least 3 of them will develop the disease?

**Answer:**

1. What is the probability that between 3 and 6 of them will develop the disease?

**Answer:**

1. What is the probability that between 3 and 6 of them, inclusive, will develop the disease?

**Answer:**

1. What is the average number of pregnant women with type 2 diabetes you would expect to see in the sample?

**Answer:**

1. What is the standard deviation of the number of pregnant women with type 2 diabetes in the sample?

**Answer:**

**Question 25[9 points]:**

Consider body mass index (BMI) in a population of 60 year old males in whom BMI is normally distributed and has a mean value of 29 and a standard deviation of 6. Using STATA:

1. What is the probability that a randomly selected male of age 60 will have a BMI of 30?

**Answer:**

1. What is the probability that a randomly selected male of age 60 will have a BMI of less than 30?

**Answer:**

1. What is the probability that a randomly selected male of age 60 will have a BMI of at most 30?

**Answer:**

1. What is the probability that a randomly selected male of age 60 will have a BMI of at least 30?

**Answer:**

1. What is the probability that a randomly selected male of age 60 will have a BMI more than 30?

**Answer:**

1. What is the probability that a randomly selected male of age 60 will have a BMI between 23 and 33?

**Answer:**

1. Assume the following BMI categories for males:

Below 20=Underweight

20 – 25=Healthy

25 – 30=Overweight

30 and Above=Obese

What is the probability that a randomly selected male of age 60 is Healthy?

**Answer:**

1. What BMI is exceeded by 25% of all BMIs (i.e. what BMI places a male of age 60 on the top 25 percentile of all BMIs)?

**Answer:**

1. What is the probability that a randomly selected male of age 60 will have a BMI of less than zero? Comment on the normally distributed assumption in this example?

**Answer:**

**Question 26 [10 points]:**

The mean blood cholesterol concentration of a large population of adult males (50-60 years old) is 200 mg/dl with a standard deviation of 20 mg/dl. Suppose we randomly sampled 100 adult males from this population, then

1. Could one find the probability that a randomly selected adult male (50-60 years old) have blood cholesterol concentration between 180 and 210 mg/dl? Why or why not?

**Answer:**

1. What is the mean of the sampling distribution of the sample mean ?

**Answer:**

1. What is the standard deviation of the sampling distribution of the sample mean ?

**Answer:**

1. Could you use the central limit theorem to determine the sampling distribution of the sample mean ? Why or why not? What is the name of that distribution if it’s determined?

**Answer:**

1. What is the chance that the sample mean blood cholesterol concentration will exceed 225 mg/dl?

**Answer:**

1. What mean values of blood cholesterol concentration cuts off the lower 10% of the sampling distribution?

**Answer:**

1. If we select repeated random samples of size 30 from this population, what proportion of the samples will have a blood cholesterol concentration larger than 225 mg/dl?

**Answer:**

1. Does the standard deviation of the sampling distribution of change as n gets smaller (from n=100 to n=30)? If yes, then how so?

**Answer:**

1. Does the mean of the sampling distribution of change as n gets smaller (from n=100 to n=30)? If yes, then how so?

**Answer:**

1. Suppose that we are also interested in the mean of the skewed serum uric acid levels. Compared to the relatively symmetrically distributed blood cholesterol concentration, do you think a larger or smaller sample size would be required to apply the CLT to the sampling distribution of the sample mean of serum uric acid levels?

**Answer:**

**Question 27 [6 points]:**

Nearly 18 of every 100 U.S. adults aged 18 years or older (17.8%) currently smoke cigarettes. Suppose you are conducting study about lung cancer and you obtain a random sample of 1000 adults.

1. What is the average number of smokers you would expect to see in the sample?

**Answer:**

1. What is the standard deviation of the number of smokers in the sample?

**Answer:**

1. What is the probability that at least 200 of the sampled adults are smokers?

**Answer:**

1. Could you approximate the binomial distribution with the normal distribution? Are the conditions for such approximation satisfied?

**Answer:**

1. Compute the probability that at least 200 of the sampled adults are smokers using the normal approximation?

**Answer:**

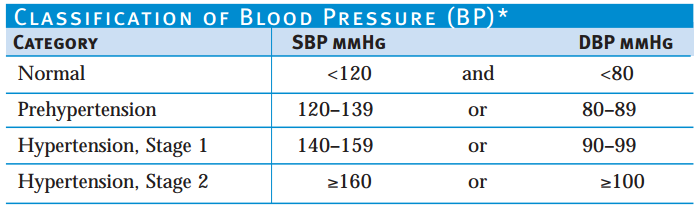
1. Are the answers in (e) and (c) similar? Explain?

**Answer:**

**Question 28[12 points]:**

Consider the following hypertension dataset (<http://www.mathalpha.com/PH-538/hypertensionfall16.dta>) [Data set 2 on the course webpage]. This dataset is courtesy of Dr Waldon Garris, University of Virginia School of Medicine. Dr Garriss collected the data in a pilot study during his work in the Dominican Republic in 1997. The subjects are persons who came to medical clinics in several villages, for a variety of complaints. Data on gender, age, systolic and diastolic blood pressure were collected.

1. Create a new variable called ***bp\_cat*** according to the following classification for blood pressure and then provide the probability distribution for the newly generated categorical variable ***bp\_cat***?



Note that we use & for “and” and | for “or” in STATA. For example, for prehypertension, we use:

replace bp\_cat="Prehypertension" if ((sbp>=120 & sbp<=139) | (dbp>=80 & dbp<=89))

**Answer:**

1. Create a new variable called ***hyper*** which classifies whether a person is hypertensive or not such that ***hyper*** is labeled “D+” if ***bp\_cat*** is either Stage 1 or Stage 2 and it is labeled “D-” if ***bp\_cat*** is either Normal or Prehypertension. You could start your code by:

gen hyper="D+" if (bp\_cat=="Stage 1" |……)

What is the sample prevalence of Hypertension in the Dominican Republic in 1997 according to this pilot study?

**Answer:**

1. Create a new variable called ***age\_cat*** which classifies the age of the person according to the three categories: 18-39, 40-59 and 60+. Please provide the PMF for the newly generated categorical variable ***age\_cat***?

**Answer:**

1. Would it be acceptable to assume that the diastolic blood pressure variable follows the normal distribution? Justify your answer?

**Answer:**

1. Would it be acceptable to assume that the sample mean, of diastolic blood pressure in a random sample of 100 subjects, follows the normal distribution? Justify your answer?

**Answer:**

1. Plot the histogram of the diastolic blood pressure variable only for subjects with the ages 50 to 58 inclusive (syntax is provided). Would it be acceptable to assume that the diastolic blood pressure variable follows the normal distribution within this age group? Justify your answer?

hist dbp if (age>=50 & age <=58)

**Answer:**

1. Would the binomial distribution be an appropriate model for the ***bp\_cat*** variable? Why or why not?

**Answer:**

1. Would the binomial distribution be an appropriate model for the ***hyper*** variable? Why or why not?

**Answer:**

1. The numeric difference between your systolic and diastolic blood pressure is called your pulse pressure (PP). Use the **gen** command in STATA to create the new variable PP and then make a histogram for the new generated variable.

**Answer**

1. Create a categorical PP variable and call it ***PP\_cat*** such that ***PP\_cat*** is “T+” if PP is larger or equal than 48 and it’s “T-” if PP is less than 48. Please provide the PMF for the new generated categorical variable?

**Answer**

1. Let’s use ***PP\_cat*** as a diagnostic test for hypertension such that a person with ***PP\_cat*** =“T+” is called hypertensive while a person with ***PP\_cat*** =“T-” isn’t. Please provide a contingency table for the two categorical variables ***hyper*** and ***PP\_cat*** (make sure that T+/T- are the rows and D+/D- are the columns)?

**Answer**

1. What are the PPV, NPV, Accuracy, Area under the ROC curve, sensitivity and specificity of the proposed diagnostic test? Is it a good test? Run the following syntax to answer this question (don’t forget to provide the output and note that you might need to slightly change the provided syntax depending on your previous labeling for bp\_cat)

gen hyper\_n=1 if (bp\_cat=="Stage 1" | bp\_cat=="Stage 2")

replace hyper\_n=0 if (bp\_cat=="Normal" | bp\_cat=="Prehypertension")

gen PP\_cat\_n=1 if PP>=48

replace PP\_cat\_n=0 if PP<48

quietly logistic hyper\_n i.PP\_cat\_n

lroc

estat classification

**Answer**

**Question 29 [3 points]:**

1. The distribution of systolic blood pressure for female diabetics between the ages 30 and 34 has unknown mean and a known standard deviation of σ=11.8 mm Hg. A random sample of 30 women is selected from this population. The sample mean systolic blood pressure is . Calculate the 95% confidence interval for µ, the true mean systolic blood pressure? Interpret this confidence interval? Please use STATA to compute the interval.

**Answer**

1. The distribution of systolic blood pressure for female diabetics between the ages 30 and 34 has unknown mean and standard deviation. A random sample of 30 women is selected from this population. The sample mean systolic blood pressure is and the sample standard deviation is . Calculate the 95% confidence interval for µ, the true mean systolic blood pressure? Interpret this confidence interval? Please use STATA to compute the interval.

**Answer**

1. In a simple random sample of 1200 adults in NM, the proportion with diabetes was found to be 0.115 (or 11.5%). Construct the 95% confident for the true proportion of adults in NM with diabetes and interpret it? Please use STATA to compute the interval.

**Answer**

**Question 30[4 points]:**

Using the hypertension dataset in question 28, please answer the following questions:

1. Construct the 99% confidence interval for the true mean systolic blood pressure? Is it appropriate to use the z-distribution to compute this interval? Why or why not?

**Answer**

1. Construct the 99% confidence interval for the true mean systolic blood pressure for females? Is it appropriate to use the t-distribution to compute this interval? Why or why not?

**Answer**

1. Construct the 90% confidence interval for the true proportion of hypertension?

**Answer**

1. Construct the 90% confidence interval for the true proportion of hypertension for females?

**Answer**