Statistics 541: Take home final exam, Fall 2006

Please support your answers by showing all of your work - this includes the TF and multiple choice questions.

1. A study was conducted here at UW to investigate hypertension in infants. Systolic blood pressure, birthweight (oz), and age (days) were measured for 16 infants. The data is shown below.

i	Birthweight (oz)	Age (days)	Systolic BP (mmHg)
1	135	3	89
2	130	4	90
3	100	3	83
4	105	2	77
5	130	4	92
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88

Suppose we assume that the relationship between systolic blood pressure (y) (SBP), birthweight (x_1), and age (x_2) is given by:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

(a) (5 pts). Is the coefficient of birthweight significantly different than zero at significance level $\alpha = 0.05$?

(b) (3 pts). What is the predicted average SBP of a baby with birthweight 8 lb (128 oz) measured at 3 days of life ?

(c) (2 pts). The predicted average SBP of infants with birthweight 6 lb
(96 oz) measured at 4 days of life is 89.106. How much higher is SBP expected to be for infants of the same weight that are one day older ?

(d) (3 pts). How far is the first infant shown in the table from the predicted average for infants of the same weight and age ? Here, distance is measured by the residual.

2. Consider two genetic mutations, G1 and G2. In a particular group of cancer patients, the probability of observing G1 is 0.6; the probability of observing G2 is 0.8; the probability that both mutations occur is 0.4.

(a) (2 pts). Are G1 and G2 disjoint ? Why or why not ?

(b) (2 pts). Are G1 and G2 independent ? Why or why not ?

(c) (2 pts). Let Y bet the number of genetic mutations present in a patient (assuming that G1 and G2 are the only possible mutations). Find the probability that Y=0.

3. The underlined portions of the following statements are either TRUE or FALSE. Indicate which answer, TRUE or FALSE, you feel is more appropriate and give a justification for it.

(a) (5 pts). Let T be distributed according to a t-distribution with 4 degrees of freedom and let Z be a standard normal. $\underline{P(T < 1.533) < P(Z < 1.533).}$

(b) (5pts). Dr. Jones performs a paired two-sample t-test (on a properly designed experiment) on data from 10 pairs. Dr. Jones tests the null hypothesis that the mean difference is "0" versus the two-sided alternative and obtains a t-value of 1.72 which results in $0.10 < \text{p val} < 0.20$. Dr. Jones argues that the original plan had been to use 20

pairs in this study but that last minute budget constraints allowed for only 10. Dr. Jones further argues that had the 20 pairs been used and had the same values for the mean and standard deviation been obtained, the t-value would have been 2.43 ($= 1.72 \times \sqrt{2}$) resulting in a p-value between 0.02 and 0.05. You find Dr. Jones' reasoning persuasive.

4. The weight of snacks in a medium-size bag is stated on the package to be 10 ounces. The amount that the packaging machine puts in these bags is believed to be normally distributed with a mean of 10.2 ounces and standard deviation 0.12 ounces. What is the probability a randomly chosen bag will be underweight, i.e. weigh less than 10 ounces?

A. 0.25

B. 0.0475

C. 0.9525

D. 0.5

5. Using the information in question 4, what is the probability that the sample mean weight of 4 randomly chosen bags is below 10 ounces?

A. 0.9525

B. 0.0475

C. 0.0004

D. 0.9996

6. A large random sample (n = 200) will be taken from a population with a mean 10 and standard deviation 8. The population is not normally distributed. Which of the following is a true statement.

A. The approximate probability the sample mean will be less than 11 can be found because of the Central Limit Theorem.

B. The approximate probability the sample mean will be less than 11 can be found using the Normal Tables.

C. The approximate probability the sample mean will be less than 11 cannot be found because the margin of error will be too large.

D. The approximate probability the sample mean will be less than 11 cannot be found because the sample mean is skewed when the population is not normal.

7. A 95% confidence interval for a population mean is (1, 6). The margin of error is

A. 1.96

B. 2.5

C. 3.5

D. 5

8. Below weight patients were given a special nutritional supplement to see if it would promote weight gain. The researchers reported that the 77 patients studied gained an average of 56 ounces and that a 95% confidence interval for the mean weight gain this supplement produces is (45, 67). Which of the following is a correct interpretation of these results?

A. 95% of the patients studied gained between 45 and 67 ounces.

B. We are 95% confident a patient given this supplement will gain between 45 and 67 ounces.

C. The true mean weight gain of a patient given this supplement will be between 45 and 67 ounces 95% of the time.

D. We are 95% confident the true mean weight gain of patients given this supplement lies between 45 and 67 ounces.

9. Which of the following will not result in a decrease in the width of a confidence interval ?

A. An increase of the sample size.

B. A decrease in the type I error specified.

C. A decrease in the standard error.

D. A decrease in σ.

10. A doctor monitoring the level of phosphorus in the blood of a dialysis patient took 6 measurements and recorded a sample mean of $\bar{x} = 5.4$ milligrams of phosphorus per deciliter of blood (mg/dl). Assuming the level of phosphorus varies normally with standard deviation $\sigma = 0.9$ mg/dl, the 90% confidence interval for the mean blood phosphorus level is:

A. 5.4 ± 1.48

B. 5.4 ± 0.9

C. 5.4 ± 0.6

D. 5.4 ± 0.37

11. Suppose for a particular measurement the standard deviation is $\sigma = 10$ milligrams. How many measurements must be made so that the standard deviation of $\bar{x}$ equals 5?

A. 2

B. 1

C. 3

D. 4

12. You have taken a simple random sample and constructed a 95% confidence interval for a population mean μ. The probability the interval contains μ is

A. 0.50

B. 0.05

C. 0 or 1 but we do not know which.

D. 0.95

13. The relationship between coffee drinking and myocardial infarction was studied in non-smoking women aged 30-49. This retrospective study evaluated 519 women and coffee drinking status and MI status were recorded. Data shown below.

Patients	$\geq$ 5 cups	$<$ 5 cups	Totals
MI	14	75	89
No MI	49	381	430
Totals	63	456	519

(a) (5 pts). Compute the odds-ratio.

(b) (5 pts). Would you reject the null hypothesis of no association between coffee drinking and MI at the 5% significance level? Explain.

14. (4 pts). Suppose that the random variable X has a normal distribution with mean equal to μ and a known variance of $\sigma^2 = 1$. A random sample of size n is taken. The

confidence interval $(\bar{X} - 1/\sqrt{n}, \bar{X} + 1/\sqrt{n})$ is a _____ % confidence interval.

15. (20pts). The scientific community acknowledges that global warming (GW) is a fact and that C02 contributes largely to GW. Are there other potential causes ? Write a 1 page summary of the current opinions regarding the causes of global warming. Which do you find most convincing and why ? Did an *Inconvenient Truth* identify likely causes, other than C02 ? Was the film's message generally convincing ? Why or why not ? Please cite the main references that you use following the summary on a separate page.

16. (20pts). Microarrays allow for the simultaneous measurement of thousands of mRNA transcript levels. In Kendziorski *et al.*, *PNAS*, 2005, we published a study that in part compared control rats (condition A) with rats treated with retinoic-X-receptor (RXR) (condition B), a potential target for some cancer therapies. Download the data set from the course website. You will see that there are 15, 923 genes measured in 24 rats (12 in condition A and 12 in condition B). Conduct gene specific t-tests assuming unequal variances to obtain a single p-value for each gene (I suggest you use Excel). Report the gene IDs and corresponding p-values for the genes with the 20 smallest p-values. What would the Bonferroni corrected value of α be if a single test's type I error is set at $\alpha = 0.05$? How many genes do you find to be differentially expressed between condition A and condition B ? Is the Bonferroni correction useful here ?