STATISTICS 641 - Exam 3

Total time is 120 minutes. The exam is available for 48 hours window, starting from noon (CST) May 01, 2014 to 11:59 am (CST) May 03, 2014.

Instructions

1. The exam is for two hours.
2. Students are allowed to bring four pages of cheat sheets (front&back, front&back, front&back and front&back). 3. Students are allowed to use calculator (need to use a calculator that can at least do normal probability calculation)

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

STATISTICS 641 - Exam #3

The following table gives the racial characteristics of 326 individuals convicted of homicide and whether or not they received the death penalty. Social scientists were interested in the relationship between defendants's race and death penalty. A possible confounding variable is the race of the homicide victim.

Defendant's race	Death Penalty Yes	No
White	19	141
Black	17	149

Let p_w and p_b be the probability of receiving death penalty when the defendant is a White and Black, respectively. Answer the next 4 questions.

1. The null hypothesis for testing the dependence between death penalty and defendant's race is

   (a) $p_w = p_b$ •
   (b) $p_w \neq p_b$
   (c) $p_w < p_b$
   (d) $p_w \geq p_b$

2. The Z test statistic for testing the above hypothesis is

   (a) 0.925
   (b) 0.017
   (c) 0.471 •
   (d) 0.654

3. For testing the above hypothesis one may also use

   (a) paired t-test
   (b) χ^2 test •
   (c) two independent sample t test
   (d) F test

4. A reasonable estimate of the probability that a White defendant receives a death penalty is

   (a) 17/166
   (b) 19/160 •
   (c) 19/151
   (d) 11/151

In a hypothestical problem suppose that we test $H_0 : \mu \leq 10$ versus $H_a : \mu > 10$ at the 5% level of significance. Suppose that the observations are normally distributed with $\sigma = 1$. Also, given that you reject H_0 if $\overline{X} > 1.96/\sqrt{n} + 10$, where $\overline{X}$ denotes the average of n random observations from that sample. Answer the next 3 questions.

5. What is the probability of rejecting H_0 when the true value of μ is 11 and $n = 9$.

   (a) 0.4819
   (b) 0.1684
   (c) 0.9986
   (d) 0.8508•

6. Determine the sample sample size to have power 0.95 when the true value of μ is 11.

   (a) 15
   (b) 9

(c) 11

(d) 13•

7. Suppose that $\sigma = 2$ but we would still reject H_0 when $\overline{X} > 1.96/\sqrt{n} + 10$. What is the probability of Type-I error?

   (a) 0.0419

   (b) 0.0865

   (c) 0.5817

   (d) 0.1635•

8. Suppose that we want to test equality of two population means. It is also known that the two populations are normal with the same variance. Which of the following statements is true?

   (a) Shapiro-Wilk's test will be the most powerful in this scenario.

   (b) The t-test at the 5% level will involve with less probability of type-I error than the Wilcoxon test at the 5% level.

   (c) The Wilcoxon test is more powerful than the t-test.

   (d) The Wilcoxon test will likely to be less powerful than the t-test. •

9. A survey is conducted of 16 year old students from inner city public schools and suburban public schools to compare the proportion who had experimented with illegal drug. The study uses

   (a) independent samples •

   (b) matched pair design

10. For testing equality of variances of the two populations we may use

    (a) t-test

    (b) χ^2 test

    (c) F test •

    (d) Fisher's exact test

    Suppose that p_G and $p_{\overline{G}}$ denote the probability of developing breast or ovarian cancer among the carrier of BRCA gene and non-carrier of BRCA gene, respectively. Consider the following table and answer the next 2 questions.

	Carrier status	
Disease	Yes	No
Cancer	45	400
No cancer	50	700

11. The odds of the disease among the BRCA carrier is

    (a) $p_G/(1-p_G)$ •

    (b) $\log\{p_G/(1-p_G)\}$

    (c) $\log(p_G)$

    (d) $\log\{(1-p_G)/p_{\overline{G}}\}$

12. The estimate of the odds ratio is

    (a) 0.454

    (b) 0.211

    (c) 0.514

    (d) 1.575 •

13. To test equality of two odds ratios one may use

    (a) Agresti-Coull confidence interval

(b) F-test

(c) Fisher's exact test

(d) Breslow test •

14. Suppose that $X_1, \cdots, X_n$ are highly positively correlated with a Normal(μ, σ^2) distribution. A 95% confidence interval for μ was constructed using the formula $\overline{X} \pm t_{0.025,n-1} s/\sqrt{n}$. The true coverage probability of this confidence interval will be

(a) 0.95

(b) more than 0.95

(c) much less than 0.95 •

(d) may be greater or less than 0.95

15. Let $X_1, \cdots, X_n$ be iid observations from a population having pdf f. The researcher wants a 95% confidence interval on σ^2, the population variance. The researcher does know that the distribution is right skewed. What is the best way of constructing confidence interval for a sample of size $n = 25$.

(a) χ^2 distribution based

(b) t distribution based

(c) using bootstrap method •

(d) predictive distribution based interval

16. Mantel-Haenszel estimator is used to estimate the common

(a) success probability

(b) odds ratio •

(c) population variance

(d) population mean

17. A researcher is attempting to estimate the average number of miles between recharging of the batteries for a battery operated car. There are several potential estimators of the average. The best approach for selecting an estimator would be to

(a) always select the unbiased estimator because its average value equals the true value of the parameter

(b) select the estimator with the smallest variance because then the estimator would have the least change from sample to sample

(c) select the estimator with the smallest average squared distance from the parameter •

(d) select the estimator with the smallest bias

(e) all the above are appropriate answers

18. The Anderson-Darling (AD) GOF statistic is prefered to the Kolmogorov-Smirnov (KS) GOF statistic for testing the goodness-of-fit of a continuous pdf because

(a) AD is a more modern procedure.

(b) AD has a more accurate p-value than does KS.

(c) AD is less likely to falsely declare that a distribution does not fit the collected data.

(d) AD is more likely to declare that a distribution function does not fit the edf, especially in the tails of the distribution. •

(e) All of the above are true.

19. A pivot $g(x, \theta)$ in constructing confidence interval is a quantity whose

(a) distribution is free from θ. •

(b) distribution may depend on θ.

(c) distribution must be Normal$(0, 1)$.

(d) tail probabilities must be a function of θ.

20. Suppose that based on a given data you obtain a 95% CI for the population proportion p, and that is $(0.224, 0.358)$. Which of the following statements is correct?

    (a) $\text{pr}(0.224 < p < 0.358) = 0.95$

    (b) If we repeatedly construct the confidence interval using the same formula and by drawing random sample from this population with the same size, then on average, 95% of those intervals will contain p, and $(0.224, 0.358)$ is one such interval. •

    (c) Although we are not sure if $(0.224, 0.358)$ contains p, we are 5% confident that it contains p

    (d) 95% of the time the true value will be around 0.291

21. Determine a lower bound on the time to failure of a device having Exponential failure times such that we are 95% confident that at least 90% of the devices will have failure times greater than the lower bound. Use n=30. Here area above $\chi^2_{r,k}$ under the χ^2 distribution with k degrees of freedom is r, for any $0 < r < 1$ and $k > 0$.

    (a) $-\overline{T}\{60\log(0.9)/\chi^2_{0.05,60}\}$ •

    (b) $-\overline{T}\{30\log(0.9)/\chi^2_{0.05,30}\}$

    (c) $-\overline{T}\{30\log(0.95)/\chi^2_{0.10,30}\}$

    (d) $-\overline{T}\{60\log(0.9)/\chi^2_{0.01,30}\}$

22. A study was designed to compare the mean yield of a genetically engineered variety of broccoli to the most widely cultivated variety. In 100 one acre plots of land, the genetically engineered variety of broccoli was raised and the annual broccoli yields were recorded: $B_1,\ B_2,\ \ldots,\ B_{100}$. From USDA records, the broccoli yields per acre of 103,259 farms were recorded: $Y_1,\ Y_2,\ \ldots,\ Y_{103259}$. The sample means $\bar{B}$ and $\bar{Y}$ are computed from the two samples. Which of the following statements is true about the bias of $\bar{B}$ and $\bar{Y}$ as estimators of μ_B and μ_Y, respectively?

    (a) $\bar{B}$ and $\bar{Y}$ have the same positive bias

    (b) $\bar{B}$ has a larger bias than $\bar{Y}$

    (c) $\bar{B}$ has a smaller bias than $\bar{Y}$

    (d) the estimator having largest bias depends on the shape of the population pdf's

    (e) none of the above statements are true •

23. Which one of the following statements is true?

    (a) The sampling distribution of the MLE, $\hat{\theta}$, of a parameter θ will have approximately a normal distribution provided n is large enough.

    (b) The bootstrap procedure for estimating the sampling distribution of a statistic provides an accurate portrayal of the distribution provided we run the bootstrap procedure a very large number of times.

    (c) If the population pdf $f(y;\theta)$ is symmetric about θ, then the sample mean $\bar{Y}$ is a better estimator of θ than is the sample median, $\widehat{\widehat{\theta}}$.

    (d) The Kaplan-Meier estimator is an estimator of the survival function when the data contain right censored observations. •

24. A kinesiology researcher is studying the stress placed on older (50-70 years) long distance runners. In order to diagnosis potential problems in this group of runners, it is necessary to establish a range of values for their resting pulse rate such that with a very high degree of confidence this range would contain 95% of the resting pulse rates for this group of runners.

    (a) The computed interval would be a confidence interval.

    (b) The computed interval would be a tolerance interval. •

    (c) The computed interval would be a prediction interval.

    (d) The computed interval would be a natural interval.

    (e) None of the above would be appropriate.

25. A large corporation is evaluating two different suppliers of a raw material. The corporation's engineers obtain a random sample of 20 units from each of the suppliers and determine a crucial product characteristic from the samples: $X_1, \ldots, X_{20}$ and $Y_1, \ldots, Y_{20}$. Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be the median for each supplier's product characteristic. It is desired to test $H_o : \tilde{\mu}_1 = \tilde{\mu}_2$ versus $H_1 : \tilde{\mu}_1 \neq \tilde{\mu}_2$. An evaluation of the two data sets reveals the following:

- The width of the box in the box plot for supplier 1 is much wider than the width for supplier 2.
- The normal reference plots have the data values very close to a straight line for both data sets.
- The two data sets are independent.

The preferred test statistic is

(a) Wilcoxon Rank Sum test

(b) Wilcoxon signed rank test

(c) Pooled t-test

(d) Separate variance t-test •

(e) Sign test