2015 National Convention — Open Statistics

You have 60 minutes to complete this 30-question test. Do not get bogged down by any one question; some are significantly more difficult than the others. The answer choice E) NOTA denotes that "None of These Answers" is correct. Remember to provide exact answers unless otherwise noted. Good luck!

For questions 1-3, use the following information: Outraged at the ever-increasing cost of taking his many dates to the movies, the Great One collected the prices of movie tickets in 70 randomly selected theaters in the US. The sample mean is $8.20 and the standard deviation is 0.6. Let μ be the population mean (i.e., the average price of a movie ticket in the US).

1. The Great One constructs a 93% confidence interval for μ and finds it to be [8.070, 8.330]. What is the meaning of such a statement?
A) If the procedure were repeated many times, 93% of the sample means would lie between 8.070 and 8.330.
B) 93% of the population of theaters charges between 8.070 and 8.330.
C) The probability that the average price of a movie ticket in the US is between 8.070 and 8.330 is 93%.
D) If the procedure were repeated many times, 93% of the resulting intervals would contain the population mean movie price.
E) NOTA

2. To the nearest hundredth, what is the smallest value μ_0 such that a test at level 7% for $\mu = \mu_0$ versus $\mu \neq \mu_0$ would fail to reject the null?
A) 8.06 B) 8.07 C) 8.33 D) 8.34 E) NOTA

3. In an interview, the CEO of Regal Cinemas said that the average price of movie tickets in the US is now $8. The Great One, after looking sadly at his empty wallet, is convinced the price is higher. Conduct a test, giving the p-value (to the nearest thousandth) and your conclusion.
A) p = 0.074, fail to reject the null hypothesis
B) p = 0.026, reject the null at 5% significance
C) p = 0.006, reject the null at all reasonable significance levels
D) p = 0.003, reject the null at all reasonable significance levels
E) NOTA

For questions 4-5, let $\lambda > 0$ be a fixed parameter. We consider a continuous random variable *X* with probability density given by $f(x) = ce^{-\lambda x}$ if $x \geq 0$, and $f(x) = 0$ otherwise.

4. What is the probability that *X* is positive?
A) $\frac{1}{\lambda}$ B) $\frac{1}{e}$ C) $\frac{c^2}{\lambda}$ D) $\frac{e}{c^2}$ E) NOTA

5. Assuming that $c = \lambda$, what is the area under the given curve?
A) $\frac{e}{4}$ B) 0.50 C) 1 D) e^2 E) NOTA

6. For a previous year's national convention, Mr. Softy wrote a statistics test that was much too easy (never fear, Mr. Softy was not invited to write this year). Which of the following could logically represent the mean, median, and mode of the results for Mr. Softy's test?
A) 110, 150, 120 B) 140, 130, 110 C) 130, 130, 130 D) 120, 140, 150 E) NOTA

7. A Type II error occurs when we…
A) Fail to reject an incorrect null hypothesis.
B) Reject an incorrect null hypothesis in favor of the alternative.
C) Reject a correct null hypothesis in favor of the alternative.
D) Do not collect a large enough sample size.
E) NOTA

8. Starcraft 2 is a video game where two players play against each other. Each one has first to choose a race among three possible choices: Zerg, Protoss or Terran. In a certain Startcraft 2 tournament with thousands of players, each race is equally represented. We want to verify that each match-up is represented in the correct proportion. To do this we are given the following data: 300 games were randomly selected among which 35 games were Terran vs. Terran, 66 Terran vs. Protoss, 74 Terran vs. Zerg, 28 Protoss vs. Protoss, 59 Protoss vs. Zerg, and 38 Zerg vs. Zerg. Compute a p-value for this test.
A) $p > 0.3$ B) $0.1 < p < 0.3$ C) $0.05 < p < 0.1$ D) $p < 0.05$ E) NOTA

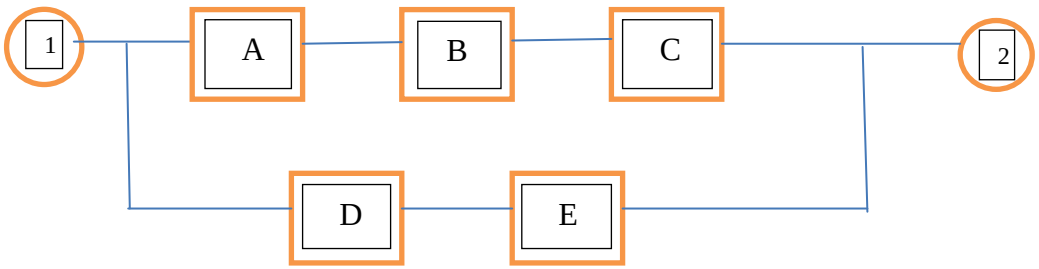
For questions 9-11, let X denote the monthly milk production (in pounds) for a randomly selected US cow between January 2004 and December 2009 (included). We know that $X \sim N(1426, \sigma)$ with an unknown standard deviation σ. Moreover, it is known that during this period, exactly 75% of the cows produced at least 1377 pounds per month. Round the z-score to two decimal places.

9. Calculate the variance of X, to the nearest tenth.
A) 73.1 B) 5,321.4 C) 5,343.6 D) 5,348.6 E) NOTA

10. What is the probability (to 2 decimals) that a randomly selected cow produces between 1400 and 1500 pounds monthly? To simplify your calculations, please round the standard deviation of X to 2 decimal places throughout the problem.
A) 0.36 B) 0.48 C) 0.50 D) 0.84 E) NOTA

11. A farmer owns 200 cows during this period and sells all of his milk to a local yogurt factory. Denote $X_1, \ldots, X_{200}$ the monthly production of each cow and assume that these random variables are independent and identically distributed. The terms of the contract with the factory stipulate that the farmer should provide x_0 pounds of milk per month 95% of the time. If x_0 must be a whole number, what is the largest x_0 that the farmer can agree to?
A) 282,535 B) 283,172 C) 283,498 D) 285,200 E) NOTA

12. Consider the circuit below. Denoting by A the event on which component A works properly, B the event on which component B works properly, etc…



We know that each component has the same probability p of functioning.
To the nearest hundredth, what is the probability that a current starting at 1 could flow to 2, given that $p = 0.6$?
A) 0.08 B) 0.50 C) 0.58 D) 0.63 E) NOTA

For questions 13-14, consider the following: A website would like to have a formula to estimate the probability of click of a visitor on the website's ads using some information about him. We denote by Y the response variable corresponding to the probability of a click, and X_1, X_2, X_3 the three covariates corresponding to age, numerical affinity with the website, and sex.

13. Assume that the covariates "age" and "numerical affinity with the website" are very highly correlated. What is the maximum number of covariates that should be included in the final regression model?
A) 0 B) 1 C) 2 D) 3 E) NOTA

14. Since the probability of clicks is likely to be very small, what would be a reasonable transformation for the dataset in order to have a meaningful linear regression?
A) $\sqrt{Y}$ B) $\log(X)$, $\log(Y)$ C) $\ln Y$ D) e^Y E) NOTA

15. The reversal of the direction of a comparison when data from several groups are combined to form a single group is known as:
A. The placebo effect
B. Simpson's Paradox
C. The Hawthorn Effect
D. Bayes' Rule

For questions 16-17, use the following information: One percent (1%) of a finance company loans are defaulted (not completely repaid). The company routinely runs credit checks on all loan applicants. It finds that 30% of defaulted loans went to poor risks, 40% to fair risks, and 30% to good risks. Of the non-defaulted loans, 10% went to poor risks, 40% to fair risks, and 50% to good risks.

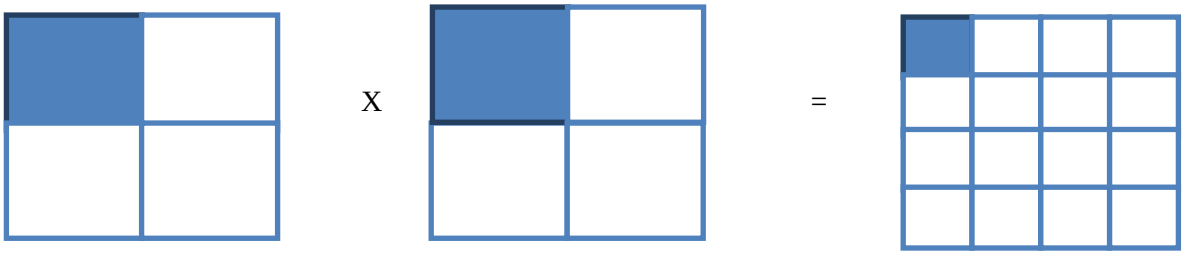
16. To the nearest thousandth, what is the probability that a poor-risk loan is defaulted?
A) 0.029 B) 0.003 C) 0.010 D) 0.102 E) NOTA

17. To the nearest thousandth, what is the probability that a loan is rated good?
A) 0.102 B) 0.396 C) 0.495 D) 0.498 E) NOTA

18. How many natural numbers lie in the set that represents the compliment of the intersection of the events $\{X \geq 3\}$ and $\{X \leq 5\}$?
A) 0 B) 1 C) 2 D) 3 E) NOTA

19. Suppose Jeremy's daily commute to class has a standard deviation of 4 minutes. What is the uncertainty (to the nearest hundredth) in his total commute time over the 5-day workweek, assuming that the commute time on any given day is independent of the times on other days of that week?
A) 20.00 B) 16.28 C) 8.94 D) 8.00 E) NOTA

20. Which of the choices is best represented by the picture below?



A) Disjoint Events B) Conditional Probability C) Bayes' Rule
D) Independent Events E) NOTA

21. Which of the following is TRUE about correlation?
A) It is a resistant measure.
B) Transforming the dependent variable by a linear operation does not change its value, but transforming by a non-linear operation (such as log or square-root) does change its value.
C) A correlation of -.8 indicates a stronger relationship between the two variables than a correlation of .7.
D) A correlation of zero indicates no relationship between the two variables.
E) NOTA

For questions 22-23: Francois Hollande has been elected as the new French president. 30 people, selected at random, were asked to estimate the likelihood (probability) that he will be successful at dealing with the euro crisis. The resulting sample mean is 0.56 with a standard deviation of 0.05.

22. We want to know if the general belief in France is that Francois Hollande has a change higher than ½ to deal successfully with the euro crisis. Conduct a hypothesis test for this problem. Give the range that includes the p-value for this test.
A) $p > 0.10$ B) $0.05 < p < 0.10$ C) $0.01 < p < 0.05$ D) $p < 0.01$ E) NOTA

23. Another set of 33 people were selected at random and asked the same question, assuming that Nicolas Sarkozy was still president. The resulting sample mean is 0.53 with a standard deviation of 0.08. We want to know if the general belief in France is that Hollande is more likely to find a way out of the euro crisis than Sarkozy. Conduct a hypothesis test for the problem, and use the conservative estimate for degrees of freedom. Under which range does the p-value you calculated fall?
A) $p > 0.10$ B) $0.05 < p < 0.10$ C) $0.01 < p < 0.05$ D) $p < 0.01$ E) NOTA

For questions 24-25: Assume that the Gator Transit shuttles arrive at the Florida football stadium exactly at times 8:00, 8:15, 8:30, etc. You arrive at a time that is uniformly distributed between 8:00 and 8:30.

24. What is the probability that you wait less than 5 minutes for a shuttle?
A) 1/6 B) 1/3 C)1/2 D) 2/3 E) NOTA

25. What is your expected waiting time, in minutes?
A) 5 B) 7.5 C) 11.5 D) 15 E) NOTA

26. Mr. Pie likes playing with the number pi, which, as you probably know, begins 3.14159265… Mr. Pie thinks there may be some sort of magical significance hidden in the digits listed above. In order to determine the meaning of it all, he decides to take each of the 9 digits as an individual data point and make a box plot of the data set. To make such a plot, he draws 5 vertical lines, representing the members of the 5-number summary. What is the ratio of the distance between the first and fifth lines to the distance between the second and fourth lines?
A) 1/2 B) 1 C) 17/9 D) 2 E) NOTA

In questions 27-29, we consider several mathematical heroes: Cantor, Galois, Gauss, Gödel, Perelman, and Poincaré. Based on their stories, Natalie believes that Cantor, Gauss, and Poincaré (respectively Galois, Gödel, and Perelman) had more or less the same productivity (in terms of the number of published papers). Moreover she believes that the first group (Cantor, Gauss, and Poincaré) was about three times more productive than the second group (Galois, Gödel, and Perelman).
To verify Natalie's belief, The Great One proposes to select at random 200 publications of the 6 mathematicians. More precisely, he collects all their publications, and selects uniformly at random 200 papers from this set. He obtains the following result:

Cantor	Galois	Gauss	Gödel	Perelman	Poincaré
52	10	66	13	12	47

27. What test do you suggest for this problem?
A) Chi-squared test for Goodness of Fit B) Chi-squared test for Independence
C) 2-sample t-test D) 2-proportion t-test E) NOTA

28. Compute a p-value (to the nearest hundredth) for this test and make your conclusion at the 5% significance level.
A) 0.07, fail to reject the null B) 0.04, reject the null
C) 0.02, reject the null D) $p < 0.01$, reject the null E) NOTA

29. When Natalie saw the table, she wanted to change her belief (that is, the null hypothesis in the hypothesis testing problem), but the Great One told her that it was not fair to do so. From a statistical point of view, how do you explain the Great One's reaction?
A) It was unwarranted – hypotheses can be changed at any point in statistical analysis
B) Natalie's actions are called "data snooping" and would introduce bias
C) Natalie's actions would inflate the probability of a Type II error
D) Natalie's actions are only acceptable when performing an ANOVA test
E) NOTA

30. When shooting at a target on a wall, suppose that the horizontal miss distance is normally distributed with mean 0 and variance 4 and is independent from the vertical miss distance, which is also normally distributed with mean 0 and variance 4. Let D denote the distance between the target and the point at which the shots lands (if x is the horizontal distance and y is the vertical distance from the target, then D $= \sqrt{x^2 + y^2}$). What is the expected value of $\frac{D^2}{4}$?
A) 1 B) 2 C) 4 D) 8 E) NOTA