

EPI 513-607 (Inferential Statistics)
 Midterm Examination, February 17, 1995

INSTRUCTIONS: Be brief and **WRITE CLEARLY**. Unless specifically asked for, complete calculations [or even complete sentences] are not needed. Answer in point form when possible. Write answers in space provided. Note that points add to >100 so you have some choice. All points obtained will be counted. If you score > 100, those points beyond 100 can be counted towards other parts of the total mark for the course.

Answers [identified via a nom-de-plume known only to you] to be handed in at/before the beginning of class on Wednesday February 22

nom-de-plume

- 1 True or false, and explain briefly [2 points each].
 - a If you add 7 to each value on a list, you add 7 to the SD.
 - b If you double each value on a list, you double the SD.
 - c If you change the sign of each value on a list, you change the sign of the SD.
 - d If you duplicate each value in a list, you leave the SD approximately unchanged.
 - e If all the values in a list are positive, you cannot have a SD which is larger than the mean.
 - f Half the values on a list are always below the mean.
 - g In a large set of measurements, the distribution of measurements follows the Gaussian curve quite closely.
 - h If two large populations have exactly the same average value of 50 and the same SD of 10, then the percentage of values between 40 and 60 must be exactly the same for both populations.
 - i An researcher has a computer file of pre-treatment White blood Counts (WBCs) for patients. They range from 2,800 to 38,600. By accident, the highest WBC gets changed to 386,000. This affects the mean but not the median and the IQR.
 - j The SD of 80 0's and 20 1's is approximately 0.4. The SD of 400000 0's and 100000 1's is also 0.4.
 - k For the men in a large U.S. sample survey (the HANES study), mean income in the different age groups increased with age until 50 or so and then gradually declined. Thus, the income of a typical man increases as he ages until 50 or so and then starts decreasing.
 - l A significance test was performed to test the null hypothesis $H_0: \mu = 2$ versus the alternative $H_a: \mu > 2$. The test statistic is $z = 1.40$. The P-value for this test is thus approximately 0.16
 - m Suppose all students in a class of 20 got the same wrong answer to a multiple choice exam question with 4 choices. To test whether the students colluded [ont triché] while the monitor was out of the room for 2 minutes, the school principal calculated the probability that a random variable Y with a Binomial(20,0.25) distribution would be 20. He did this by first calculating $\mu=20(.25) = 5$ and $SD=\sqrt{\frac{0.25 \times 0.75}{20}}$. He then calculated $\text{Prob}[Z \frac{20-\mu}{SD}]$ and, finding that the P-value was very small, he concluded that the students had "almost certainly" colluded. List two problems (1) with the main calculation error and (2) an even bigger logical error in inference; ignore the issue of continuity corrections and the accuracy of the Gaussian approximation.

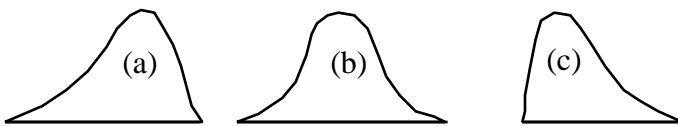
2 [6 points]

The diameter of red blood cells in healthy adults has a Gaussian distribution with mean 7.5 microns and standard deviation 0.3 microns. Approximately what fraction of red blood cells have a diameter between 7.1 and 7.6 microns?

Would the standard deviation of red blood cells be about the same in a bigger person (with more red blood cells) as in a smaller person (with fewer red cells)? Why/why not? What *general* lesson does this example give about the relation between the SD of individual observations and the number of observations? [1 sentence]

3. [5 points]

As part of a survey, a large company asked 1000 of its employees how far they commute to work each day (round trip). The average round trip commute distance was 18 Km, with an SD of 25 Km. Would a rough sketch of the histogram for the data look more like (a) or (b) or (c)? Or is there a mistake somewhere? Explain your answer.



4 [9 points]

An athlete suspected of having used steroids is given two tests that operate independently of each other. Test A has probability 0.9 of being positive if steroids have been used. Test B has probability 0.8 of being positive if steroids have been used. If steroids have been used, what is the probability that both tests are negative? *A tree may help*

- (a) 0.02 (b) 0.72 (c) 0.30 (d) 0.28 (e) none of the above

Suppose the two tests are indeed negative. What can now be said about whether the athlete has used steroids? [*one sentence*]

Use this example to explain why the alpha used in a statistical test cannot be used in the same way following statistical tests as positive and negative predictive values are following medical tests. [*one sentence*]

5 [5 points]

In a study of the effects of acid rain, a random sample of 100 trees from a particular forest are examined. Forty percent of these show some signs of damage. Which of the following statements are correct?

- (a) 40% is a parameter
(b) 40% is a statistic
(c) 40% of all trees in this forest show signs of damage
(d) more than 40% of the trees in this forest show some signs of damage
(e) less than 40% of the trees in this forest show some signs of damage

6 [6 points]

In which of the following would X not have a Binomial distribution? Why?

- a. X = number of women in different random samples of size 20 from the 1990 directory of statisticians.
- b. X = number of occasions, out of a randomly selected sample of 100 occasions during the year, in which you were indoors. (One might use this design to estimate what proportion of time you spend indoors)
- c. X = number of months of the year in which it snows in Berlin.

7 [4 points]

A significance test gives a P-value of .04. From this we can... [indicate True/False for each]

- (a) reject H_0 at the $\alpha = .01$ level
- (b) reject H_0 at the $\alpha = .05$ level
- (c) say that the probability that H_0 is false is .04. [be careful!]
- (d) say that the probability that H_0 is true is .04. [be careful!]

8 [9 points]

A health department serves 50,000 households. As part of a survey, a simple random sample of 400 of these households are surveyed. The average number of adults in the sample households is 2.35, and the SD is 1.1.

- a Sketch a possible frequency distribution showing the variability in the number of adults per household [don't spend a lot of time on trial and error getting the distribution to match the mean and SD exactly; only the general shape is required]
- b If possible, find an approximate 95%-confidence interval for the average number of adults in all 50,000 households. If this isn't possible, explain why not.
- c All adults in the 400 sample households are interviewed. This makes 940 people. On the average, the sample people watched 4.2 hours of television the Sunday before the survey, and the SD was 2.1 hours. If possible, find an approximate 95%-confidence interval for the average number of hours spent watching television by all adults in the 50,000 households on that Sunday. If this isn't possible, explain why not.

9. [6 points]

In a simple random sample ($n=225$) of all institutions of higher learning in the U.S., the average enrollment was 3,700, with an SD of 6,000. A histogram for the enrollments was plotted and did not follow the normal curve. However, the average enrollment at all institutions in the U.S. was estimated to be around 3,700 ($SE = 400$). Say whether each of the following statements are true or false, and explain why.

(a) It is estimated that 95% of the institutions of higher learning in the U.S. enroll between $3,700-800 = 2,900$ and $3,700 + 800 = 4,500$ students.

(b) An approximate 95%-confidence interval for the average enrollment of all institutions runs from 2,900 to 4,500.

(c) If someone takes a simple random sample of 225 institutions and goes two SEs either way from the average enrollment of the 225 sample schools, there is about a 95% chance that this interval will cover the average enrollment of all schools.

10 [6 points]

An investigator at a large university is interested in the effect of exercise in maintaining mental ability. He decides to study the faculty members aged 40 to 50, looking separately at two groups: the ones who exercise regularly and the ones who don't. There are large numbers in each group, so he takes a simple random sample of 32 from each group, for detailed study. One of the things he does is to administer an IQ test to the sample people, with the following results:

| | regular exercise | no regular exercise |
|---------------|------------------|---------------------|
| sample size | 32 | 32 |
| average score | 132 | 120 |
| SD of scores | 16 | 16 |

The difference between the averages is "highly statistically significant". The investigator concludes that exercise does indeed help to maintain mental ability among the faculty members aged 40 to 50 at his university.

a State the null and alternative hypotheses, calculate the p-value and verify the statement about the difference being "highly statistically significant".

b Is the investigator's conclusion justified? Why/why not?

11 [6 points]

An investigator wants to show that first-born children score higher on vocabulary tests than second-borns. She will use the WISC vocabulary test (after standardizing for age, children in general have a mean of 30 and a SD of 10 on this test). She considers two study designs:

- i In a school district find a number of families with both a 1st-born and a 2nd-born enrolled in elementary school.
 - ii From schools in the district, take a sample of 1st-born and a sample of 2nd born children enrolled in elementary school.
- a List 1 statistical and 1 practical advantage of each approach.
 - b For the design you prefer, what would you recommend as a statistical test of the hypothesis?
 - c For design ii, and assuming she tells you that a difference of 3 points on the standardized test would be important, determine an appropriate sample size. (The sample size for design i involves the correlation between the scores of 1st and 2nd born children)

12 [5 points]

A colony of laboratory mice consisted of several hundred animals. Their average weight was about 30 grams, with an SD of about 5 grams. As part of an experiment, graduate students were instructed to choose 25 animals haphazardly, without any definite method. The average weight of these animals turned out to be around 33 grams. Is choosing animals haphazardly the same as drawing them at random? Discuss briefly, carefully formulating the null hypothesis, and computing Z and P. (There is no need to formulate an alternative hypothesis)

13 [10 points]

"We studied the involvement of naturally occurring odours in guiding the baby to the nipple. One breast of each participating mother was washed immediately after delivery. The newborn infant was placed prone between the breasts. Of 30 infants, 22 spontaneously selected the unwashed breast. The washing procedure had no effect on breast temperature. We concluded that the infants responded to olfactory differences between the washed and unwashed breasts". Abstract from "Does the newborn baby find the nipple by smell?" Varendo H et al., Lancet 1994 344: 989-90.

- a. What is the parameter at issue here?
- b. State the implied null and alternative hypotheses concerning this parameter.
- c. What is the statistic reported?
- d. What 'reference distribution' should be used to test the hypothesis?
- e. Given the sample size involved, how would you calculate the p-value? [actual calculation not required]

14 [8 points]

Refer to the article "Inhibition of oxidation of low-density lipoprotein with red wine" by Kondo K et al. Lancet 344 page 1152 October 22, 1994

- a Calculate the SD and the variance of the 10 lag times at Day 0 and at Day 14.

- b The authors used 'error bars' of ± 1 SE rather than \pm some larger multiple of the SE, presumably so as to ensure that the two CI's for day 0 and day 14 did not overlap. If you were going to put a 95% CI at each point, what multiple would you use?

- c If the authors calculated a CI for the mean difference as $54.7 - 49.1 \pm m \sqrt{2.6^2 + 2.2^2}$, they would find that for any multiple m bigger than about 1.65, the corresponding confidence interval included a mean difference of zero. Does this mean that the difference is not statistically significant at conventional levels of significance? How does one reconcile this with the $p < 0.01$ reported by the authors?

[Calculating the SE based as $\sqrt{s^2[1/10 + 1/10]}$ where s^2 is the weighted average of the two variances, would give the same SE in this example; so the issue is not one of separate versus pooled variances! Nor is it an issue of 1- versus 2-sided tests!]

15 [15 points]

Refer to the article "Is evidence for homeopathy reproducible" by Reilly D et al. Lancet 344 page 1601-1606 December 10, 1994 [we will use this for several statistical procedures next week]

- a Verify the sample size calculation [1st paragraph of Analysis Section, 2nd column, page 1602].

- b From the data given, verify the $p=0.003$ and the CI [-24.1 to -5.9] given in the 1st paragraph of the 1st column of page 1604.

- c Suppose you had only been given the CI. Show how to reconstruct the p-value from it.

- d From a quick reading of the report, do you have any serious concerns? Or do you generally agree with the conclusions?

16 [10 points; for those who were not challenged enough!]

A snail (Schnecke/escargot) starts out to climb a wall. During the day it moves upwards an average of 22 cm (SD 4 cm); during the night, independently of how well it does during the day, it slips back down an average of 12 cm (SD 3 cm). The forward and backward movements on one day/night are also independent of those on another day/night.

a After 16 days and 16 nights, how much vertical progress will it have made? Answer in terms of a mean and SD. Note that -- contrary to what many students in last year's exam calculated -- the SD of the total progress made is not 80 cm; show that it is in fact 20 cm.

b What is the chance that, after 16 days and 16 nights, it will have progressed at least 150 cm?

c Over and above the assumption of independence, which was 'given', did you have to make strong [and possibly unwarranted] distributional assumptions in order to answer part b? Explain carefully, giving justification.