Department of Epidemiology and Biostatistics, McGill University

EPI 513-607 (Inferential Statistics)
Midterm Examination, May 31, 1994

INSTRUCTIONS

Be brief and W R I T E C L E A R L Y .

Unless specifically asked for, complete calculations [or even complete sentences] are not needed. Answer in point form when possible.

Answers [identified via a nom-de-plume known only to you] to be handed in at/before the beginning of class on Tuesday June 7. We will collectively correct/grade some of the work in class.

1 True or false, and explain briefly.

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 list, the distribution of measurements follows the normal 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 The variance of the sum of two random variables is always bigger than the variance of each one.

j An researcher has a computer file of pre-treatment 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.

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 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 Prob[$Z \geq \frac{20-\mu}{SD}$] and, finding that the P-value was very small, he concluded that the students had "almost certainly" colluded. [Hint: there are several problems; concentrate on the main calculation error and also on the bigger problem of a possible logical error in inference; ignore the issue of continuity corrections and the accuracy of the Gaussian approximation]

2 Refer to the letter to the BMJ from a left-handed medical statistician concerning a serious bias in the comparison of ages at death of left-handed and right-handed persons. The same point would apply, even more dramatically, if we were to compare age at death of persons who went through 'the new math' curriculum in elementary school with age at death of those who had the 'old math' curriculum (the 'new math' curriculum was introduced into western countries at various stages in the 1960's and 1970's). It would also apply to a comparison, via the obituary columns of the medical journals, of age-at-death of radiologists (theirs is a long established specialty] and emergency-medicine specialists (an emerging specialty).

To demonstrate that you understand Peto's point, ...

a construct a realistic 2-way table describing the age distributions in these two types of persons {l/r or, if you prefer, newer/older} in a 1994 population [or as Peto looks at it, the prevalences of these two types of person as a function of age]. To keep it simple, limit yourself to one gender.

b Apply the same age-specific death rates to the two types of persons [if you wish, you can use the death rates derived from 1990 Quebec mortality data given in page 3 of the material on M&M $4.1; you can still make the same point if you use fewer age categories to reduce the amount of arithmetic].

c Then, for each of the two types, calculate the average age-at death of those that die in the next several years. How big a difference do you get in the "average age at death" of the two types of persons? Comment.

3 In the eighteenth century, yellow fever was treated by bleeding the patient. One eminent physician of the time, Dr. Benjamin Rush, wrote:

> *I began by drawing a small quantity at a time. The appearance of the blood and its effects upon the system satisfied me of its safety and efficacy. Never before did I experience such sublime joy as I now felt in contemplating the success of my remedies.... The reader will not wonder when I add a short extract from my notebook, dated 10th September, 1793]. "Thank God, of the one hundred patients, whom I visited, or prescribed for, this day, I have lost none."*

Explain some of the design problems in Rush's study.

4 A snail 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?

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.

5 A overview of randomized clinical trials of antiplatelet therapy as prophylaxis against deep venous thrombosis [BMJ on 22 Jan. 1994] found the following:

Category of trial	% odds reduction (SD)
general surgery	37% (8)
traumatic orthopaedic surgery	31% (13)
elective orthopaedic surgery	49% (11)

a Is 37% a statistic or a parameter? Why?

b Does each SD refer to (i) variation of individuals or (ii) sampling variation associated with the estimate? Explain your reasoning.

c Use the SD of each estimate to argue that the apparent heterogeneity in the percent reductions, i.e. the spread from 31% to 49%, could simply reflect random variation alone [differences among three estimates are more difficult than we have learned to dealt with, so for simplicity, concentrate on the difference of two estimates]

Since we have neither a statistical nor a biologic basis for assuming different size effects for different types of patients, in the spirit of Occam's razor, we can construct one overall estimate from the three. One way to do this is take a simple average of the three reductions, giving each estimate a weight of 1/3 i.e. (37+31+49)/3 = 39%.

d If we create this equal-weighted average, the uncertainty {SD} associated with it should be smaller than the SD of the components. Calculate the SD for
$\frac{1}{3}\text{estimate}_1 + \frac{1}{3}\text{estimate}_2 + \frac{1}{3}\text{estimate}_3$ [M&M §4.3 & exercises 4.64–4.66 should help]

Since we have three estimates with different degrees of uncertainty, it makes sense to calculate an average of them which gives more weight to the individual estimates with smaller SD's. It can be shown mathematically that the weighted average with the lowest SD is the one with weights that are inversely proportional to the individual variances. In our example here, this would lead to weights proportional to $\frac{1}{64}$, $\frac{1}{169}$ and $\frac{1}{121}$ respectively, or an overall estimate of $0.52\text{estimate}_1 + 0.20\text{estimate}_2 + 0.28\text{estimate}_3$. This gives a weighted average of just over 39%. [the fact that the two methods give almost the same answer is a coincidence in this example; it doesn't happen generally]

e As stated, this information-weighted average has a lower uncertainty {SD} associated with it than a simple equally-weighted average. Calculate the SD for
$0.52\text{estimate}_1 + 0.20\text{estimate}_2 + 0.28\text{estimate}_3$; compare it with the SD in d above.

The overview reported an estimate of 42% (SD 17) for high risk medical patients.

f Combine the single estimate for surgical patients and the 42% for medical patients. Your overall estimate should be closer to 39 than to 42%. Why? Calculate its SD. Why does the 'averaging' of the two estimates not diminish the SD very much?

Comments: 1. It might be preferable (in terms of more accurate CI's) to combine the estimates on a log scale, where the sampling variation should be more Gaussian; 2. the above method of combining estimates (and particularly of calculating SD's, or SE's if you prefer) is not appropriate if there is definite heterogeneity in the estimates (also, the average has much less meaning).

6 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; if you can show one which comes within 0.1 of the mean and 0.2 of the SD, that's good enough]

b If possible, find an approximate 95%-confidence interval for the average number of adults in all 50,000 households, and from it an approximate 95%-confidence interval for the total 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.

7 New laser altimeters can measure elevation to within a few inches, without bias, and with no trend or pattern to the measurements. As part of an experiment, 25 readings were made on the elevation of a mountain peak. Their mean was 81,411 inches, and their SD was 30 inches. Fill in the blanks in part (a), then say whether each of (b-e) is true or false; explain your answers briefly. (You may assume Gaussian variation of the measurements, with no bias.)

a The elevation of the mountain peak is estimated as _____ . There is approximately a _____ % chance that we are over-estimating or under-estimating the true elevation by more than 6 inches.

b 81,411 ± 12 inches is a 95%-confidence interval for the average of the 25 readings.

c 81,411 ± 12 inches is a 95%-confidence interval for the elevation of the mountain peak.

d A large majority of the 25 readings were in the range 81,411 ± 12 inches.

e The elevation of the mountain peak is the statistic here; the 81,411 is a parameter.

8 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 author's conclusion justified? Why/why not?

9 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 2-child 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.

c For the design you prefer, what would you recommend as a statistical test of the hypothesis?

b For the design you prefer, and assuming she tells you that a difference of 3 points on the standardized test would be important, determine an appropriate sample size. If you don't have sufficient information to make the determination, explain to her exactly what she needs to provide you before you can determine the sample size.

10 Consider a RCT that led to the recommendation of lumpectomy and radiation as an equally effective but less disfiguring alternative to mastectomy in treating breast cancer. In the original study there were three treatment groups: total mastectomy (n = 590), lumpectomy (n = 636), and lumpectomy and irradiation (n =629). At the end of the follow-up period (average 81 months), the numbers alive with no evidence of disease were: total mastectomy 373 (63.2%), lumpectomy 371 (58.3%) and lumpectomy and irradiation 412 (65.5%). [I haven't checked these numbers; they, and questions a-c that follow are taken from an article in Chance News[1]]

a Calculate a margin of error associated with each of the percentages alive with no evidence of disease. Likewise, calculate a margin of error associated with the difference of the first and third percentages. State your level of confidence that the errors in the estimates are no more than what you have calculated. What are the most important assumptions are you making in calculating these limits of error?

b What would be the effect on these margins of error if the data on a random 19% of the study subjects were removed? Carry out the calculations.

c Suppose that some women enrolled were technically ineligible for the study, although the randomized assignment and follow-up were properly carried out in an unbiased way. The research group said that a new analysis, with the data on 19% of the patients removed, shows that the study's original published conclusions remain valid. But a government spokesman remarked that removing 19% of the sample diminished the statistical power of the study. What does this latter statement mean?

d You were asked to participate in deciding the sample size for a new two-arm study to revisit the question of total mastectomy versus lumpectomy and irradiation. Given the intense public interest in the new trial, the oncologists in the research group ask you, as the most statistically articulate, to provide technically accurate interpretations of the 3 Greek symbols (α, β, Δ) in the sample size formula that would be understandable to journalists and educated non-experts in statistics (or for that matter in clinical trials). You might also be interviewed by a local television station. Prepare such an interpretation, limiting yourself to 200 words [or 150 seconds of 'sound bites'] in total.

e If you are female, what value of Δ do you think should be used? If you are male, ask some (statistically?) significant female in your life (who hasn't taken a course in statistics) what value of Δ should be used. How would you word your question to her?

---

[1] Prepared by J. Laurie Snell as part of the CHANCE Course Project supported by the National Science Foundation and the New England Consortium for Undergraduate Science Education. Current and previous issues of CHANCE News can be found on the internet via gopher to: chance.dartmouth.edu.

11 Refer to the article "Hair concentrations of nicotine and cotinine in women and their newborn infants by Eliopoulos et al (JAMA 1994; 271:621-623).

a The authors state that the sample size was chosen to detect twofold more cotinine in infants of passive smokers than infants of non-smokers [last paragraph of Subjects and Methods]. This "twofold more cotinine", roughly speaking, corresponds to a difference of 0.3 between means in the $\log_{10}$(concentration) scale, a scale on which the observations are more nearly [but still not quite] Gaussian than in the concentration scale. Suppose that their pre-study information was that the between-infant SDs on this $\log_{10}$ cotinine scale would be approximately 0.4 for each of the two groups being compared. Assuming they were going to recruit equal numbers of passive smoking and nonsmoking mothers, and with the alpha and power they mention, how many of each would be required?

b If cotinine measurements were Gaussian on the $\log_{10}$ scale, would they be Gaussian on the ln i.e. $\log_e$ scale? Note that $\log_{10}$ (cotinine) = $0.4343\log_e$ (cotinine) = $0.4343\ln$(cotinine).

c For the 36 active smoking women, the mean number of cigarettes used daily was 11.4. What was the SD? Why would this between-woman SD be of little use in describing the pattern of between-women variation in reported consumption [stated to have varied from 1 to 40]?

d In the last sentence of the first paragraph of Results, what do (i) the statement that "r=.75" and (ii) the word "significant" mean?

e Why do you think "there was no correlation between the daily number of cigarettes reported by the mothers and either maternal or neonatal concentrations of nicotine or cotinine"?

f Put the statement "P<0.001" [after the r=.49 at top of third column] into words that these parents would understand. Don't use the circular explanation that because P<0.001, it is "significant".

g "Maternal concentrations of nicotine were <u>invariably</u> higher than neonatal levels (P<0.001)" [next sentence]. Since this certainly isn't the case for <u>all</u> <u>94</u> mother pairs in Figure 1, the authors must be referring only to the n=36 pairs where the mother smoked. The authors don't say in their statistical methods section what test they used to calculate this p-value [they only refer to tests for 'between groups']. What 2 tests of hypotheses that are covered in M&M Ch 7 were available to them? Exactly what hypothesis does each one test?

h In plain words, what is meant by the phrase "concentrations of cotinine did not differ significantly between mothers and infants"?

i The primary endpoint of interest was stated to be infants' hair concentration of cotinine, and the sample size calculation concentrated on the passive smoking versus non smoking mothers. Mean {SEM] concentrations of cotinine in infants of passive smoking and nonsmoking mothers were 0.60[0.15] and 0.26[0.04] respectively. The authors say that these concentrations were significantly different.
Just from the numerical summaries {mean[SEM]} they provide, can you can perform a statistical test to verify this? Do you feel comfortable carrying out this test? Why/Why not? If not, and if you had access to the detailed data, what other options would you propose?

j The Figure legend doesn't say, but what do the error bars in Fig 3 represent? Would you have used something else? Why/Why not?