Department of Epidemiology and Biostatistics
McGill University

EPI 513-607 (Inferential Statistics)
Final Examination
December 8-10, 1993

INSTRUCTIONS

The answers are to be written in the spaces provided.

When writing, be brief and WRITE CLEARLY.

Unless specifically asked for, complete calculations are not needed. To avoid writing out formulae, just indicate which table or formula would be appropriate and give a reference; explain where one obtains each of the components of the formula.

The points (indicated in **bold** beside each question) add up to 230. The exam will be marked out of 180, so one point deserves one minute of effort.

____
your ID number or *nom-de-plume*

4 If the government cuts the salaries of all employees by a flat amount of $70 per month, what does this do to the average monthly salary? *{Reduce it by $70}* to the SD?*{Nothing}*

If, instead, it cuts all salaries by 3%, what happens to the average and the SD? *{Both Reduced by 3%}*

6 A list has 100 numbers. Each number is either 1, 2 or 3.

(a) the average is 2 and the SD is 0. What is the list? *100 2's*

(b) the SD is 1. What is the list? *50 0's 50 1's*

(c) can the SD be bigger than 1? *No*

3 Is the SD of the age of 1st year university students closer to 1 month, 1 year or 5 years? Explain your reasoning *1 month is too little, even if all were say 18 years old; 5 years is too much.*

4 In a random sample of 400 workers from a working population of 20,000 persons, the average distance they travel to work was 19 Km, with a SD of 20.0 Km. Find an approximate 95% CI for the average distance that all workers in the town travel to work. If this isn't possible, explain why not.

xbar ± t_{399} • SEM or 19 ± 1.95 • 20 / √400 or 19 ± 1.96(1) or 17.08 to 20.92

n=400 & Central Limit theorem => skewness of x's not an issue in sampling variability of xbar

4 4 histograms are shown below for the following variables (in a study of a small town):

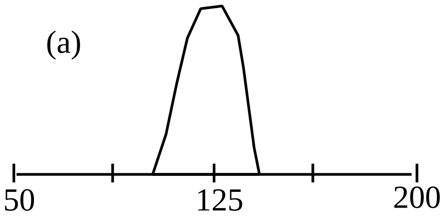
b<==>(i) heights of all members of households where both parents are less than 25 years old

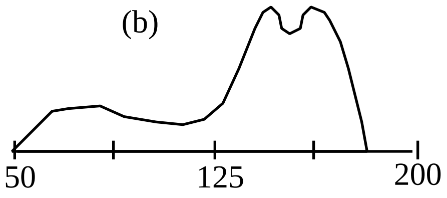
c<==>*(*ii) heights of married couples

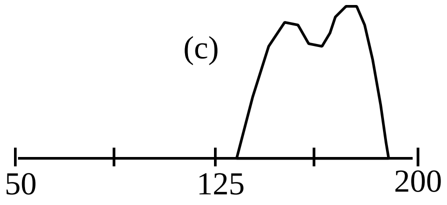
d<==>*(*iii) heights of all people

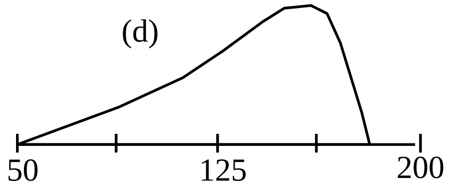
a<==>*(*iv) heights of automobiles (cars only)

Match the variables to the diagrams (all heights are in cm's). Explain your reasoning.









5 As genetic theory shows, there is very close to a 50:50 chance that both children in a 2-child family will be of the same sex. Here are two possibilities:

(i) 10 couples have 2 children each. In 8 or more of these families, it will turn out that both children are of the same sex.

(ii) 20 couples have 2 children each. In 16 or more of these families, it will turn out that both children are of the same sex.

Which possibility is more likely, and why?
# same sex is Binomial with π*=0.5 and n=10 or 20. (i) Prob (# ≥ 8 | n = 10) = 0.044+0.010+0.001*
which is much bigger than (ii) Prob (# ≥ 16 | n = 20) = 0.005+0.001+0.000+0.000+0.000
Use Binomial table or formula or Table on Sign Test

- 25 measurements are made of the speed of light. Their average is 300,007 and their SD is 10 Km/sec.

2 Fill in the blanks: The speed of light is estimated to be *300,007;*

a 95% CI is approximately *300,007±* t_{24} *• 10/√25 or 300,007± 2.064 • 2 or ± 4.128*

True or False? explain your answers

1 The measurements differ from 300,007 by an average of 10 or so.
True*, if you condider that the SD as a kind oi 'average' deviation' about xbar and forget that you divided by 24 and not 25. [actually* Sd^2 *is an average* $deviation^2$ *]*

1 The average of the 25 measurements differs from 300,007 by 2 or so. ***False*** *(it differs by zero; 2 is SEM, and so is a measure of variability of xbar around μ, or in this case c ... physicists should have given the speed of light a Greek letter to denote that it is a parameter)*

1 If a 26th measurement were made, it would differ from the speed of light by 2 or so.. ***False***, *since the 2 refers to the SEM, the variability of xbar (not of the individual measurements). Also, the measurements would be distributed around c (the speed of light) and not around 300,007 and the average deviation around c is ~ 10 (10 is our best estimate of σ and thats what σ means).*

1 A 95% CI for the speed of light is 300,007 ± 4. ***True***, *if don't fuss about decimals.*

1 A 95% CI for the average of the 25 measurements is 300,007 ± 4. ***False***; *it doesn't make sense to talk about a CI for xbar. CI is for a parameter (c here), not for a statistic.*

1 Approximately 95% of measurements are within a range of 20 Km/sec. ***True***, *if use 4 SD's (2 on each side) and assume that the SD of 10 is a reasonably accurate estimate of σ.*

1 If another 25 measurements are made, there is a 95% chance that their average will be in the range 300,007 ± 4 Km/sec. ***False***; *there is a good chance they will be in the interval c ± 4 Km/sec.*

8 Fill in the blanks in the following letter to the editor arising from an article on how to estimate the frequency of dizygotic twinning:

I am puzzled by a calculation in X's article. In it he states that a single ovulation with intercourse will result in a live birth with probability 1/4 and therefore a double ovulation will result in dizygotic twins with probability 1/16. He concludes that the double ovulation frequency must be 16 times the observed dizygotic twinning rate among births.
This conclusion is incorrect since a large fraction of the double ovulations will not lead to a birth at all. A double ovulation will yield zero births with probability 3/4 x 3/4 = **9/16**, one birth with probability 1/4 x 3/4 + 3/4 x 1/4 = **6/14**, and two births with probability 1/4 x 1/4 = **1/16**. Hence the probability that a double ovulation will yield twins, conditional on there being any birth at all, is **1/16 / { 1/16 + 6/14}** i.e. **1/7.** Therefore X should conclude that the double ovulation frequency is **7** times the observed twinning rate. X should use a tree diagram to see this more clearly.
*Note: whereas the 1/7 calculation is correct if we assume p(survival of egg) = 1/4, there is a logical error in the sentence " Therefore X should conclude that the double ovulation frequency is **7** times the observed twinning rate". We can say that (with intercourse) D double ovulations produce on average D/16 twin births and 6D/16 singleton births; S single ovulations produce an average of S/4 singleton births. Thus, the oberved twinning rate r_T (calculated as the number of twin births divided by the total number of singleton plus twin births) is an estimate of D/16 divided by {D/16 + 6D/16 + S/4 }. If one reverses the algebra and solves for the D:S ratio, it works out to*

$$D/S = 4r_T / (1 - 7r_T),\ \text{so with say } r_T = 1/100,\ \text{we get } D/S = 0.04 / (1 - 0.07) = 4/93$$

equivalently, using probabilities rather than odds,

$$D / (D+S) = 4r_T /(1 - 3r_T),\ \text{or } 0.04 / 0.97 = 0.041.$$

The further correspondence to the letter above (by Atkinson DE in Nature 322 780 1986) corrects the error and gives a more general formula that incorporates the probability per egg of a live birth.

- 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.

5 After 16 days and 16 nights, how much vertical progress will it have made?
Total vertical progress $VP = VP_{D1} + VP_{N1} + VP_{D2} + VP_{N2} + \ldots + VP_{D16} + VP_{N16}$.

$E(VP) = (22 - 12) + (22 - 12) + \ldots + (22 - 12) = 160$

$Var(VP) = 4^2 + 3^2 + 4^2 + 3^2 + ... + 4^2 + 3^2 = 400$ so $SD(VP) = \sqrt{400} = 20$.

One could be 95% sure that the snail was somewhere between 120 and 200 cm from base.

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

Assuming VP is $N(\mu=160, SD=20)$, then $prob(VP \geq 150) = prob\{Z \geq (150–160)/20\}$

$= prob\{Z \geq –0.5\} = 0.69$

2 Did you have to make strong distributional assumptions in order to answer the previous part?

Assumption of Gaussian for VP is reasonable because even if individual components not Gaussian, by CLT the sum of 32 independent components will be a lot closer to Gaussian

3 The inner planets (Mercury, Venus) are the ones closer to the sun than the Earth is. The other planets are farther away. The masses of the planets (data from 1962) are shown below, with the mass of the Earth taken as 1.

Mercury	Venus	[Earth]	Mars	Jupiter	Saturn	Uranus	Neptune	Pluto
0.05	0.81	[1.00]	0.11	318	95	15	17	0.8

The masses of the inner planets average 0.43 while the masses of the outer planets average 74. Is this difference statistically significant? Or does the question make sense?

Not really, since there is no sampling involved (at least as far as we know!) If were to test, would use a non-parametric test such as Rank Sum test (arguing that variances were very unequal, no way to check Normality, etc . In fact, with n=2 and n=6, there are no configurations that are significant at the 0.05 level, let alone the ranks of 1 and 4 (sum = 5) that the inner 2 would get among the n=8.

5 One large course has 800 students, broken down into section meetings with 25 students each. The section meetings are led by teaching assistants. On the final exam, the class average is 65, with an SD of 15. However, in one section, the average is only 61. The assistant is accused of bad teaching. He argues this way

> If you took 25 students at random from the class, there is a pretty good chance that they would average below 61 on the final. That's what happened to me – chance variation.

Outline carefully the steps and calculations involved in this argument (do not complete the calculations)

General principle of invoking (random) sampling variation as first possible explanation is good; if 800 split up into 25-student sections at random, averages for different sections would vary around 65 with a SD of the averages somewhere in the neighbourhood of $15/\sqrt{25}$ or 3 (we call this the SEM). Because of the CLT, the variation across (among) the section averages would be close to Gaussian even if the variation across the 800 individual marks were not. From Gaussian curve, could calculate the proportion of sections that would be expected to have an average below 61, as $Prob[Z < (61–65)/3] = Prob[Z< –1.33] = 0.09$ so expect 32x0.09 or 3 sections to have averages this low or lower, even if all that were operating were random sampling from a single universe with mean 65 i.e. lower results in some section should have happened anyway.

- Refer to the attached material on "Math Survey" from NY Times. You need only consider the diagram entitled "How the States Fared on Math" and the 1st paragraph of the 2nd column of text (approximately 2,500 students...)

4 Approximately what was the SD for the national data (1st row)? What assumptions are you making in order to estimate it?

<u>If individual scores were Gaussian</u> (the symmetry is a good indication they may be) the interval

325 – 215 = 110 corresponds to the interval from –1.645 to +1.645 on the Z scale i.e. 3.29 Sd's in all. So 1Sd should be about 110/3.29 or 33. Note that the SD of the 37 sample avrages does not estimate the national σ for individuals

*Incidentally, there is a question as to whether the "averages" shown in the table are really averages (**means**) or **medians**. The numbers given do not fit with the middles of the box plot plots. We suspect that what are denoted averages on the box blots are really **medians** (50%iles), especially since the 5% and 95% values are given as well.*

Assume for the next 3 parts that the data for each state are based on simple random samples of 2500 children per state.

4 (BEFORE LOOKING AT THE DATA FOR INDIVIDUAL STATES) For state i, imagine the average score μ_i for ALL CHILDREN (sampled or not) in the state. Suppose there were no, or only trivial, differences in the 37 μ's. Nevertheless, even if this were true, there would still be (sampling) variation among the averages found in the 37 samples. How much would this variation be?

Can describe using SEM = SD / √n or 33 /√2500 or 0.66 marks. Even without Gaussian-ness of individual marks, , variation of the xbars would be Gaussian (because of CLT) with mean μ and SD(xbar's) = SEM = 0.66. (The formula for SEM is general; it requires simple random random sampling but does not appeal to CLT per se). Thus the oberved averages should be only 1 or 2 points different from each other .

3 Given your answer to the previous part, and given the observed variation in the 37 estimates of state averages (from a high of approximately 290 to a low of approximately 255), what can you say about the "supposition" that there are no differences in the 37 μ's?

Instead of a variation of say 4 SEM's from one end to other, the variation in xbars is huge and not at all compatible with the variation expected from simple random sampling from states all of which had the same mean μ.

4 How would you calculate a 95% CI for the difference between the μ for New York and the μ for its neighbour, New Jersey? Do not complete the calculations.

As {\O(x,¯)$_{NY}$ – \O(x,¯)$_{NJ}$} ± 1.96 SE{\O(x,¯)$_{NY}$ – \O(x,¯)$_{NJ}$} = {\O(x,¯)$_{NY}$ – \O(x,¯)$_{NJ}$} ± 1.96 \R({s$^2_{NY}$ / 2500 + s$^2_{NJ}$ / 2500})

In fact, the data for each state are based on random samples of 2500 children in about 100 public schools.

3 How does this fact affect the estimates of the μ's or of differences between them ?

*SE's would be bigger, CI's wider etc, relecting fact that each 2500 is a cluster sample and that there is likely to be more resemblance between the marks of children in the same school than of children from different schools i.e. we don't have 2500 completely **independent** observations per state.*

3 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 50 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	50	50
average score	132	121
SD of scores	15	15

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?

Is this conclusion justified? Yes ___ No *X*

Check one, and say why. *Inference is that exercise <u>helps</u> maintain IQ ; no justification for this at this late stage ... would need to look at this prospectively and avoid questions as to what IQ was like before beginning regular exercise, and what else is different between the two groups etc.. It really needs an experiment to sort out. It may be that persons with a higher IQ <u>think</u> its better to exercise!* ***Remember that a highly significant difference just says that there is a nonzero difference [that sampling variation alone would not produce differences as big as those observed] but it doesn't explain what caused it.***

- 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:

a In a school district find a number of 2-child families with both a 1st-born and a 2nd-born enrolled in elementary school.

b From schools in the district, take a sample of 1st-born and a sample of 2nd born children enrolled in elementary school.

4 List 1 statistical/practical advantage of each approach.

• *design (a) will remove a lot of 'noise' [due to variations in scores between children of different families that are more to do with genetics and environments] and be more efficient in terms of sample size*

• *design (b) is easier to carry out (would need to go to more schools for (a); also (b) allows direct matching on [ie elimination of effects of] age, whereas (a) will require synthetic matching [standardization of tests by age]*

3 For the design you prefer, how would you guide her on sample size considerations?

If prefer (a): matched pair analysis [paired t]. Need estimate of σ_d*, the variation across pairs in their within-pair differences [if cannot get this directly, can think of* $VAR(d) = Var(x_1) + Var(x_2) - 2Cov(x_1,x_2) = 2\ Var(x)\ [\ 1 - \rho\ ]$ *where* ρ *is the correlation between members of the same family with respect to their age-adjusted scores.* $Var(x) = 10^2 = 100$*. I expect that estimates of* ρ *can be found in the literature; if not I would put it conservatively at say 0.4 or 0.5 [its probably higher]. One also needs to decide on a 'delta' that would be considered substantial or meaningful, as well as of course specifying the alpha and beta levels. Then see sample size formula for a 1-sample test (of paired differences)*

If prefer (b): 2 independent samples [2-sample t]. Given $\sigma = 10$*, the variation across children in general with respect to their age-adjusted scores. Delta , alpha and beta levels as above. Then see sample size formula for a 2-sample t- or z-test (of differences in* μ*'s)*

3 For the design you prefer, what would you recommend as a statistical analysis? *See above*

4 If instead of a quantitative test, one were interested in whether first-born children are more likely to be left-handed than second-borns, what form would your recommended analysis take?

Analysis of proportions, either independent samples (easier to describe) for design (b) or trickier layout for matched pairs [with say 1st born as 2 rows (L/R) and 2nd born as columns (L/R).] and McNemar analysis of discordant pairs.

- A table in a publication went something like the following (denominators imputed, but information not otherwise altered).

Rates and Rate ratio for hip fracture for men 65 years of age and older in communities with fluoridated vs nonfluoridated water. [MY = man-years of observation]

Fluoridation status	Hip fractures	Rate/1000MY	Rate ratio	95%CI(Rate ratio)
Fluoridated (MY= 5475*]	19	3.47	1.41	1.00 to 1.81
Nonfluoridated [MY=14159*]	32	2.26	1.00	...

* not reported; back-calculated by JH from the numbers of cases and the rates. [Incidentally, JH is unable to figure out how 3.47 divided by 2.26 gives 1.41, so in fact it is difficult to know where the miscalculation or transcription error is. However, for the purposes of the questions in the exam, assume the ratio and the CI are correct; JH also finds it strange that, with only 51 fractures in all, the CI for a ratio is so "symmetric" but then the authors give an unusual reference, which he has not had time to check, for how they calculated the CI of the ratio from the SE's of the rates).

It drew the following letter:

> [...]. I am however puzzled by the claim of a significantly increased risk of hip fracture due to artificial fluoridation in [both] men [and women]. Although the male subjects exposed to artificial fluoridation did demonstrate a higher rate than those not exposed, the 95% CI for the ratio includes 1.00. This would seem to indicate that the increased rate is not statistically significant.

and reply:

> [...] We did have a lower confidence limit that was equal to 1.00. Our P value was equal to *0.05 (2-sided).* There is still a one in *20* chance that the data may be wrong. We chose to report our result as statistically significant: perhaps reporting the P value in this case would have been less confusing.

3 Just from their CI, fill in their P value directly

If 95% CI just 'touches' 1.0, and if using same basis for test and CI, then testing the observed ratio against the null of 1 should result in a two sided P-value of exactly 0.05.

4 Rather than getting the P value via the CI, how would one calculate it from the raw data (assume, to make it easier, that they followed 5475 and 14159 persons, all of them 70 years old, for one year each, with no losses to follow-up and no deaths, and that they found 19 and 32 hip-fractures respectively) Calculations not necessary

Test of 2 proportions, using either z test or chi-square test

4 Is the authors' interpretation of a P value correct? Can you say it better?.

It's not really correct. Are data ever wrong? are decisions wrong? The P-value is a statement about data conditional on a null hypothesis. In this case we can say that even if fluoridation had nothing to do with hip fractures, there is still a 5% chance of getting a difference as big as or bigger than we observed. It takes a lot more to be able to say that the hypothesis is wrong or right or to be able to say how sure you are that it is.

- In a comparison of the degree of pain experienced when 74 diabetic patients took blood samples from their fingertip versus their abdominal wall, the report states

  "Degrees of pain on finger versus abdominal sampling were, respectively, severe 20 vs 0, normal 31 vs 0, slight 22 vs 2, hardly any 1 vs 7 and none 0 vs 65. Some 88% found abdominal sampling completely painless."

  This is all that the authors said, but here is a 'translation':

	Severe	normal	slight	hardly at all	none
finger	20	31	22	1	0
abdomen	0	0	2	7	65

5 Suggest possible statistical tests of these data, and comment on each one

Ideally, since this is a self-paired study, we would want the pairs or ratings. Then we could carry out a signed rank test, or simply a sign test, for example (one might object to treating the rating as numerical enough to do a paired t-test). Alternatively, one could dichotomize the scale into + or – and do a McNemar test on the discordant pairs [One student # 9354587 did this and got 9 + on both, 0 – on both, 65+ on finger but – on abdominal sampling, 0 + on abdominal but – on finger sampling, and 0 + on both; so the analysis rests on the 65:0 split of discordant pairs]. If we ignore the self-pairing, we can continue to use the ordered nature of the responses by say using a rank sum test [see article on comparison of ordered responses by Moses referred to in course] or - if we are willing to put a number to each grade, a test of trends in proportions [the latter is interchanging the role of stimulus and response to 'make it fit', while the rank sum test is more natural]. Of course in this study, the data are so clearcut anyway that we don't need formal tests!

5 What proportion of all diabetic patients might expect to find abdominal sampling completely painless?

Note the word ALL. Several simply said 88% [± 0]. Students who remembered the name of the course added a binomial-based CI.

- Refer to attached excerpt from " Suicides by asphyxiation in NY city after publication of *Final Exit.* "

4 Fill in the blanks in the authors' section on statistical analysis:

"the method of suicides by methods recommended in Final Exit during the year after its publication was compared with the number that would have been expected if the publication of the book had no effect. The expected number was defined as

half the total of the two years

Those who said 'the same as the previous year' are forgetting the argument (especially under the null hypothesis) that the previous year may have been a randomly low year and the year after a randomly high one, even if the book had not been published.

3 "The binomial test was used to determine whether there was a statistically significant change". Why does the binomial come into it? (it looks like chi-square test with observed and expected numbers)

They tested a 50:50 split of the 41 suicides into the year previous and the year after.

4 For comparing the overall numbers of suicides, the authors say in a footnote that they used the normal approximation to the binomial distribution. How does one do this? (don't worry about the continuity correction, which makes only a small difference with the large numbers involved)

Testing that the total no. n split evenly (or rather randomly) over the two years, so testing if the *observed split is compatible with a Binomial split with π=0.5. Say y = no. in year after; then E(y) = nπ and Var(y)=nπ(1–π). Here π=0.5. Use y as a Gaussian variable.*

2 If someone used a chi-square test instead of the binomial (again without correction), would one get the same p-value? Why? why not?

Yes, since binomial is approx z and z^2 *= chi-square*

2 Why is there a 'continuity correction' with these kinds of data?

We are taking a discrete (integer) variable and mapping it onto a continuum (Integer ± 0.5) so we can use the Gaussian distribution.

- This letter appeared in the Lancet in 1991.

"Dr X and her colleagues' report on transverse limb-reduction defects after chorion villus sampling (CVS) at 56-66 days' gestation prompted us to examine data in the Italian Multicentre Birth Defects Registry. The possibility of an association between CVS and birth defects is a matter of such concern that a report of our preliminary findings is warranted.
From 6604 cases of malformation reported in 1988-90 we selected the 118 with a transverse limb-reduction defect, this being the one in all of Dr X's cases. The controls were the other 6486 malformed cases. 4 cases of limb-reduction and 15 controls were delivered after CVS (odds ratio 15.14, 95% CI 4.18 to 49.65; p[Fisher's exact] = 0.000305. This approach could underestimate the association, if CVS were to be associated with more than one type of defect; it is a conservative approach, appropriate to the testing of a specific hypothesis, as here...."

6 From what JH can tell, the alternative hypothesis was one-sided. For their data, list the data points (tables) that make up the 'one tail' in Fisher's exact test.

Obserrved table			Others more extreme (make CVS look worse)						
	Cases	Controls							
CVS +	4	15	5	14	6	13	....	19	0
CVS –	114	6471	113	6472	112	6473	....	99	6486

2 Give a reference for how to calculate the 2-tail area if the alternative hypothesis is 2-sided.
JH's notes on Fisher's exact test show an excerpt from Armitage and Berry where they talk about (a) doubling the probability in the 1 tail or (b) adding probabiliies in the other tail that are less than the probability of the observed table

2 Why do you think the authors used this 'exact' technique?
Probably because the expected numbers in some cells are less than 5 or some other 'warning number' that made them feel that the chi-square test is not accurate.

- "Our survey a few years ago found that many of the 41 milk samples we tested contained amounts of Vitamin D quite different from those stated on the label. In order to assess if the increased awareness of the problem has improved matters, we recently purchased 94 milk samples in several locations in the U.S.A. and Canada. Unfortunately, the overall distribution of vitamin D concentrations in these new samples is similar to what we found earlier, i.e. there has been no improvement in the fortification process. In view of the consequences of under- and over-fortification, a unified national program is needed to ensure that proper amounts of vitamin D are included in milk". [adapted from a recent report].

6 How would you summarize the raw data from the more recent surveys?

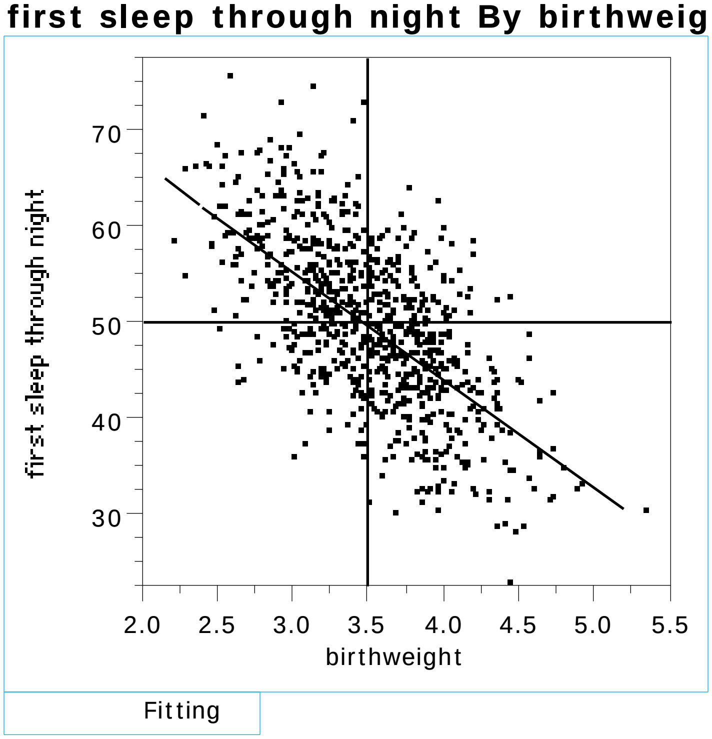
The approach should be DESCRIPTIVE, giving the magnitudes of the discrepancies. Would first ask an expert what is a serious under or over-fortification (eg <50% of stated amount or say >200% of stated amount. Then tell how many were seriously discrepant and in which direction etc. A statistical test for the mean is not enough as it doesn't describe variation in individual samples [It is false comfort to know that <u>on average</u> the milk has 100% of what the label states*

when say half of units could have 0% and half could have 200%! The emphasis here should be the same as that in the issue of agreement of measurements with each other or with some standard. A box plot would help show the raw data and the magnitudes of the discrepancies. A test of whether the mean is as advertised will be misleading without some sense of ***magnitude****.*
**Second report (Chen TC et al NEJM p 1507 vol 329 no 20, 1993) used categories <5%, 5-80%, 81-120, >120%.*

4 How would you compare the results with those of the earlier surveys?
See if discrepancies are less; maybe compare the % within 'acceptable' range.

- A study of 800 babies looked at the relationship between their weight at birth and the age at which they first slept through the night. The birth weights averaged 3.5 Kg with an SD of 0.5 Kg. The ages at which they first slept all night averaged 50 days with an SD of 10 days [JH's consultants on these data believe that both the mean and the SD for this latter variable should be much larger!]. The correlation between the two variables was –0.60.

10 Draw a rough scatterplot of what the raw data look like (you don't have to plot all 800 points).



key items:

- *axes birthweight on x-axis; age on y-axis;*
- *scales on each axis (if Guassian then most from 3.5Kg–2SD = 2.5 KG to 3.5Kg + 1Kg = 4.5Kg likewise, from 50–2(10) to 50+2(10) for age i.e. from 30 to 70 days*
- *negative correlation so low birthweight<--> later age and vice versa*
- *r = –0.6 means that the variance of y's at any x = 1 – 0.6^2 = 64% of overall variance in y i.e. SD(Y | x) = 80% of var in all y's i.e 80% of 10 or 8 days. still a fair amount of scatter in individual y's at each x*

5 Draw in the regression line. Hint: use the "centered" form: $\mu_{y|x} = \mu_y + ß (x - \mu_x)$

Line goes through the point x=xbar=3.5, y=ybar=50

slope = r • SD(y) / SD(x) = –0.6 • 10 / 0.5 = –12 days per Kg of birthweight.
so if take x = 2.5 Kg, then E(y | x=2.5) = 50 + (–12)(–1) = 62 days

5 If the birthweights were in grams rather than Kg, what would ß be? What would the correlation be? Likewise, if the age was measured in weeks, what would change?

ß is –12 days per Kg or –12/1000 i.e. –0.012 days/gram; Correlation unchanged

If age in weeks, slope is –12/7 weeks per Kg; correlation again unchanged

5 If we consider a baby that weighed 2.5 Kg at birth, what is the probability that it will sleep through the night before it is 10 weeks (70 days) old? You don't need to DO the calculation, just indicate HOW to. What distributional assumptions do you have to make?

If we knew that mean is 62 days, SD is 8 days, then can use
Prob(Z > (70 – 62) / 8) i.e. Prob(Z>1)

We are assuming Gaussian-ness of the individual variations here.

Since unsure about where mean y is for infants with x=2.5 (after all the line is just and estimate of where the mean is), would need to add this to the uncertainty via the formula

70 = 62 + z • 8 • \R(1 + \F(1,800) + \F([2.5–3.5]^2 , ∑{x - 3.5}^{2})))

i.e. get Prob(Z > $\dfrac{70-62}{\sqrt{1+\dfrac{1}{800}+\dfrac{[2.5\text{–}3.5]^2}{\sum\{x-3.5\}^2}}}$ *)). With n so large here, the second and third terms under the square root sign are negligible.*

5 If we consider all babies that weigh 2.5 Kg at birth, what is the probability that the <u>average</u> age at which they will first sleep through the night is less than 70 days?

This is a question dealing with where we think $\mu_{y|x=2.5}$ *is. Our best estimate is 62 days. The uncertainty is given by its SE i.e. by • \R(\F(1,800) + \F([2.5–3.5]^2 , ∑{x - 3.5}^{2}))) so we can use CLT and assume Gaussian uncertainty.*

- Refer to attached article "Effects of beer on breast fed infants"

4 How would you - a priori, obviously - have decided the sample size for this study?

Ask experts what would be an important reduction ∆ in amount consumed. Need some idea of the SD of a within-child difference in consumption from a four hour period in one day to a four hour period in another day. Could do a pilot study with say 10 children measured [without any intervention] on two different days and get the SD of the 10 differences. Then use sample size formula for 1 sample t-test with whatever alpha and beta decided upon.

4 Do you have a way to reconstruct the SD of the 11 within-pair differences? If yes, explain how; if not, why not?

We know t = dbar / SD(11 differences) /√11 = 2.47. From xbar1 and xbar2 we have that dbar = 193.1 – 149.5 = 43.6 so we can work back to SD(11 differences) = 43.6 • √11 / 2.47 = 59.

4 What do you think the ±18.4 and ±13.1 are? What are other possibilities and why do you tend to rule them out?

Just by their size they are too small to be SD's measuring the variation across 11 children in one session (children are not that homogeneous). Working back from the SD of 59: We know that the

SD of a difference is roughly $\sqrt{2}$ times the SD of each individual set if the 2 sets of observations are uncorrelated. Thus if SD(difference) = 59 = $\sqrt{2}$ SD(individual observations in one session) we would have SD(across individuals at one session) = 42. If there was a correlation r between the 2 sessions ie if a child who was above the average of the 11 on one session tended to be above/below the average of its group on the other session, then the SD(11 differences) = 59 would equal SD(one session) • $\sqrt{2}$ • $\sqrt{(1-r)}$. But there is no sensible r such that we could get an SD of 59 from SDs of 13 or 18.

So the 13.1 and 19.4 must be SE's or 2 SE's of the mean at each session. Since one wouldn't expect that r is very large (especially if children were all the same age), one would guess that the SD of 42 or so in a session is not that far off, and if we divide the 42 divided by $\sqrt{11}$ to get a SEM, it is not be too far from the 13 and 18 reported

4 Are you comfortable with the statistical analysis performed? List 2 other tests that were available to the authors.

One could question use of t-test with such small n=11 where we would are unable -- even from the raw data -- to check normality and would have to rely on our expert judgement.
So instead of relying on the t-test and its uncheckable assumptions, we could use the sign test or the signed rank test (nboth nonparametric)

4 In the last paragraph, why are the authors careful about their inferences?

The hypothesis is about milk production but the study measured milk consumption, and did so in only one 4 hour session for each condition. There is the question of blinding, of long term behaviour, of what the mechanism is, etc.

- Refer to attached letter entitled "Thornton Wilder's original design"

5 Suppose, before funding him, the funding agency had asked Brother Juniper to document how reproducible his ratings (and overall index) were. Suppose he had consulted you. What data would you have advised him to assemble to answer this question and what data summaries/presentations would you have recommended?

Would want to see how he would do if asked to re-rate subjects on his scale. Also, would want to see if others understand and can reproduce it. This doesn't even get into the question of the validity of the scale; the first step is usually to see if it is reproducible.

I would advise him to use Bland and Altman article ***Lancet 1986 1 307-310*** *(or CV or something quite descriptive) to present results of intra-rater and inter-rater reproducibility.*

5 Suppose the pestilence left 9 survivors for every 1 it carried off, and that his budget and other constraints only allowed him to study a total sample size (cases and controls) of only 30 "souls" ("subjects"). Why would Brother Juniper use samples of 15 and 15 rather than say 27 and 3?

If the variation in the two populations is the same, the SE of the estimated difference is proportional to \R(1/n_1 + 1/n_2) and this is minimised by having n_1=n_2; a 3:27 is not efficient.

5 Translate the last sentence of the passage from the book ("He added up ... ") into the way one might report the results today (you do not have to use a ratio).

Total(victims) = 5• Total(Survivors) or xbar(victims) = 5 • xbar(Survivors) . Of course today this would be accompanied by a t-test and a P-value. How about this answer from one student?
THE DEAD WERE 5 TIMES MORE LIKELY TO SURVIVE THAN THE SURVIVORS

5 If he submitted his report to the journal Religio-Epidemiology, the reviewers might had concerns about study design, data analysis, and interpretation. What issues of a more statistical nature might they have raised?

blinding; sample size; power; reproducibility; statistical significance; clinical significance;

5 Not all statistical tests/procedures available today were available in Brother Juniper's time. What relevant statistical tests would have been available to Brother Juniper if he were analyzing his data today? *t-test for independent samples (1910 or so); rank sum test (1940's)*