EPI 513-607 (Inferential Statistics) Midterm Examination **+ ANSWERS ,** May 14– 23, 1997

INSTRUCTIONS*: Be brief and WRITE CLEARLY. Unless specifically asked for, complete calculations [or even complete sentences] are not required. Answer in point form when possible. Write answers in space provided, or on back of sheet if necessary. Completed exam to be handed in at/before the beginning of class on Friday May 23*

Team entries welcome (maximum: 4 per team)

1 [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. Indicate with T or F which of the following statements are true and which are not.

F (a) 40% is a parameter *did not examine the entire forest*

T (b) 40% is a statistic *did examine a sample of the forest*

F (c) 40% of all trees in this forest show signs of damage *??; didn't look at all*

F (d) more than 40% of the trees in this forest show some signs of damage *ditto*

F (e) less than 40% of the trees in this forest show some signs of damage *ditto*

2 [5 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 the probability that the athlete has used steroids? *[one sentence]*

If the tests are any good (ie. are positive more often in steroid-taking persons than persons not taking steroids, then yes, our post-test probability that the athlete is taking steroids should be less than our pre-test probability (whatever it was). The pre-test probability is a function of all that we knew about the athlete pre-test (and may vary from one diagnostician to another)

Use this example to explain the limitation of a p-value in the interpretation of statistical tests. If you like, make the analogy with medical tests *[one sentence]*

$P(H|data) \neq P(data|H)$ ... the p-value is only a part of the overall interpretation, just as the predictive value of a test depends not only on the operating characteristics of the test, but also on the type of person to whom the test is applied.

3 [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.

Binomial. Some thought that $\pi(F)=\pi(M)=0.5$ was required . not so !

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)

Binomial. Its the sampling of occasions that makes the independence and same π for the 100 sampled occasions.

c. X = number of months of the year in which it snows in Montreal.

False. π(snow) nor same each month; variation from year to year in # months with snow tighter than binomial with n=12 and and say ave(π) = 0.45.

4 [4 points]

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

F (a) reject H_0 at the α = .01 level

T (b) reject H_0 at the α = .05 level

F (c) say that the probability that H_0 is false is .04.

F (d) say that the probability that H_0 is true is .04.

Remember that $P(H_0 \mid data) \neq P(data \mid H_0)$

5 [5 points]

The following is part of a table in a recent paper from the Annals of Internal Medicine on a randomized placebo-controlled trial of low-dose aspirin in patients with chronic stable angina (paper courtesy of Leslie Brailsford from a previous summer)

"Baseline Characteristics of Participants with Chronic Stable Angina in the U.S. Physicians' Health Study

Characteristic	Aspirin Group (n=119)	Placebo Group (n=102)
Mean age, years	63.6 ± 9.3	62 4 ± 8.6
Mean systolic blood pressure, mm Hg	132.5 ± 13.0	132.5 ± 14.4
Mean diastolic blood pressure, mm Hg	80.3 ± 7 8	80.2 ± 7.9
Mean cholesterol level, mmol/L	5.9 ± 1.1	5.8 ± 1.3

Plus-minus values are mean ± SE "

If you were checking this paper for typographical and other errors before it was published, would you have noticed any statistical error(s)? Explain.

Authors included Hennekens; typos were introduced inadvertently by me; but the so-called SE's given cannot be SE's; if they were, then the SD's they imply ($SD = SEM \times \sqrt{n}$) would be huge and unrealistic)

6 [8 points]

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

- Fill in the blank: The speed of light is estimated to be ... *300,007;* a 95% CI is approximately *$300{,}007 \pm t_{24} \bullet 10/\sqrt{25}$ or $300{,}007 \pm 2.064 \bullet 2$ or ± 4.128*

True or False? explain your answers

- The measurements differ from 300,007 by an average of 10 or so.
  ***True**, if you consider that the SD as a kind of 'average' deviation' about xbar and dont fuss that you divided by 24 and not 25. [actually SD^2 is an average $deviation^2$]*
- 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 ... statisticians would have given the speed of light a Greek letter to denote that it is a parameter)*
- 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 s means).*
- A 95% CI for the speed of light is 300,007 ± 4. ***True**, if don't fuss about decimals.* ***CI is FOR a parameter***
- 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.* **CI is not FOR a STATISTIC**; it is derived from statistics
- 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 s. wording "2o on either side of c " better*
- 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. MANY thought this statement was true.*

  THIS IS THE MOST COMMON WAY TO MIS-STATE WHAT A CI IS. See 101 ways to say it wrong on pages 50-53 of CoursePack get in the phr

7 [18 points] Effects of Beer on Breast-fed Infants

To the Editor.—In response to a query in JAMA about the value of beer consumption to the breast-feeding mother[] it was concluded that there was a scientific basis for the folklore that beer is a galactagogue[] [galactagogue: "favours the production of milk" -- Dorland's Medical dictionary] Beer, unlike other alcoholic beverages, increases serum prolactin levels[]. The subjects in these studies were normal men and nonlactating women however. To our knowledge, no investigation in this area focused on the lactating women and, perhaps more importantly, determined whether milk intake by breast-fed infants is enhanced when their mothers drink beer.

Recently, we demonstrated that breast-fed infants consumed significantly less milk during the 3-hour testing session in which their mothers drank a small dose of ethanol in orange juice than when the mothers drank orange juice alone.[] Using similar methods, we now report similar effects following alcoholic beer consumption. Each of 11 nursing mothers and their infants was tested on 2 days separated by 1 week. In a counter-balanced fashion, the mother drank a 0.3-g/kg dose of alcoholic beer (Miller, 4.6% vol/vol alcohol) on one testing day and an equal volume of nonalcoholic beer (Miller Sharp's, <0.5% vol/vol alcohol) on the other day. During the next 4 hours, each infant fed on demand. Milk intake was assessed by weighing the infants immediately before and after each feed, the infants' behaviors during breast-feeding were monitored by videotape, and the mothers' perceptions of their lactational performance were determined via questionnaire.

Consistent with our previous findings[], the infants consumed significantly less milk during the testing session in which their mothers drank the alcoholic beer(l49.5± 13.1 mL) than during the session in which the mothers drank the nonalcoholic beer (193.1 ± 18.4 mL, paired t *[df,* 10]=-2.47; P=.03). The mothers were apparently unaware of this difference, however. Regardless of whether they consumed alcoholic or nonalcoholic beer, most mothers believed their infants had ingested enough milk, reported that they experienced a letdown during nursing, and felt that they had milk remaining in their breasts at the end of the feeds. Analyses of the video records revealed that infants terminated approximately 70% of the feedings on each testing day. There was no significant difference for any of these variables between the two conditions.

Because milk intake and the rate of milk synthesis varies from feed to feed,[] a 23% reduction in milk intake may be difficult for women to perceive. Additionally, unlike the bottlefeeding mother who often feeds her infant in response to the amount of formula remaining in the bottle,[] the breast-feeding mother does not have an obvious means of assessing how much milk her infant consumed. Moreover, breast-feeding imposes a more active role on the infant; the infant often determines the pace and duration of the feeding and regulates the amount of milk ingested.[] These factors may explain why the folklore that beer consumption enhances lactational performance has persisted for centuries.

These findings do not imply that occasional beer consumption would decrease overall milk intake by the infant. Nor do they directly test whether beer consumption does or does not act to stimulate the amount of milk produced by the mother. However, they do suggest that such folklore should be carefully evaluated using rigorous methods.

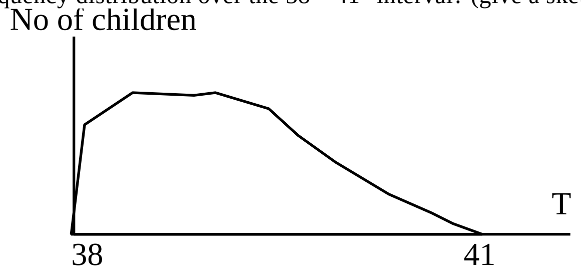
- How would you - a priori, obviously - have decided the sample size for this study?
  Ask experts what would be an important reduction Δ in amount consumed. Need some idea of the SD of a within-child difference in consumption from a four hour period in one day to a four hour period in another day. Could 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.
- 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.
- 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 appear to be 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 √2 times the SD of each individual if the 2 sets of observations are uncorrelated. Thus if SD(difference) = 59 = √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) • √2•√(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 √11 to get a SEM, it is not be too far from the 13 and 18 reported
  Incidentally, the reasoning that the SD is too big (small) for the small sample size misses the point that SD is a measure of (in this case) inter--individual variability and that while a large sample size will give a more reproducible (reliable) estimate

of σ, it is difficult to predict whether the SD in a smaller sample will be bigger (smaller) than the SD in another larger sample.

- Is the p value of 0.03 1- or 2-sided? *2-sided if we check against table*
- 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 (both nonparametric)
- In the last paragraph, why are the authors careful about their inferences?
  Issue of production vs consumption, long term vs 1--time , etc...
  finding ?? goes against 'conventional wisdom'. or 'folklore'..

8 [30 points] Paracetamol and Fever

a Entry was limited to children with temperatures between 38°C and 41°C.
Given the mean of 38.9 °C and the SD of 0.9, what can you say about the shape of the frequency distribution over the 38°- 41° interval? (give a sketch)



b *"We estimated a sample size requirement of 210 subjects completing the trial"* (Sample size — paragraph 5 of Methods)
Give the formula by which the authors estimated this (identify what numbers go with what parameters, but leave the calculations to your assistant [who has not taken a statistics course])

Eqn on R side of p73 of course pack, with $z_\alpha = 1.96$, $z_\beta = -1.645$, $\Delta=1$, $\sigma=2$
or Table on page 74 with signal to noise ratio of 1/2=0.5

c *"Student's t- test and Mann-Whitney (alias Wilcoxon) test..."* (Statistical testing — paragraph 5 of Methods)
Why did the authors use the Mann-Whitney (alias Wilcoxon) test? In light of the n's and the shape of the distribution of duration of fever, was their concern about the use of the t test justified?
Distrn. had long right tail, and so technically, the t-table isn't 100% accurate. BUT, n's are quite large and so would expect the Central Limit Theorem to apply and the t table to be quite good.

d *"The mean duration of fever..."* [paragraph 4 of Results]
Explain in a sentence, in non-technical words, the phrase "the differences were statistically non-significant"

there was not sufficient evidence in the data to reject the (null) claim that children receiving Tx and Placebo have the same mean duration of response. The observed difference in means is within the range of variation one might expect if all children received the same tx (or same placebo).

e "The 95% CI for the differences between the paracetamol and placebo groups for duration of fever was -10.0 to +7.1 h"
Explain in non-technical words what this statement says.
We are 95% confident that had we continued on to study a very large number of such children in each group, the difference in means would be in the range -10 to +7.1h. we are using a procedure that 'traps' the correct value of the parameter in 95% of applications.

f How does this CI add to what is shown in Figure 1?

Just from Fig 1, we do not know whether the reason for the 'ns' is (i) there is a great deal of incertainty about $\mu1-\mu2$ or (ii) the uncertainty is small (narrow) enough that the 'ns' can be taken as a "definitive" negative.

g How was the CI calculated?

ave duration - ave duration ± $t_{(223,95)}$ SE (ave duration - ave duration)

t close to z=1.96 and SE(diff) = $\sqrt{s^2\{1/123 + 1/102\}}$*, where* s^2 *is the pooled (weighted average) variance.*

h Before the study, the authors anticipated a SD of 2 days (48 hours) for the duration of fever. The SD of the duration of fever observed in the n=225 is not reported explicitly.

How could one reconstruct this SD from the results given [assume that the SD is the same in the two treatment groups]?

margin of error = 8.55 = t × $\sqrt{s^2\{1/123 + 1/102\}}$*, so can backcalculate to get s = 32.5*

i "Children..were more likely to be rated.as having at least a 1-category improvement in activity...." [2nd last paragraph of Results]

What tests could be used to compare the two groups? Do they all give the same answer?

z or x^2 *to compare 2 proportions (give same p-value ... 2 sided(*

rank sum test for ordered responses (uses more than 2-point scale so should be more sensitive than using just a 2-point 0-1 scale)

j *"On the basis of ...completing the trial"* [sample size considerations, first sentence of paragraph 5 of Methods]

"There were no significant differences between groups in mean duration of subsequent fever" [Abstract]

If these two statements were the ONLY information you were given about the trial, what could you conclude?

The trial was planned so that if a difference of 24 hr exists, the study had a 95% chance of finding a statistically significant difference...

The trial did not produce a statistically significant difference even though the authors gave it 'every opportunity' (95% power). If we reverse the logic, and play a bit loose with the semantics, we might say that we are reasonably sure that the difference is less than 24 h (I shouldn't really switch from P(data | H) to P(H | data)..).. safer also to use the calculated CI (-10 to +8 h) rather than going back to pre-trial arguments about what WOULD BE IMPORTANT... after all, the data have spoken! not about WOULD but about what the difference IS

9 [15 points] Melatonin and Delayed Sleep

a What sample size formula or table would you have shown the authors if they had consulted you concerning sample sizes before doing their study?

Page 63 L hand side, or table on p 64. Key is that this is a 1-sample study.

b What is it about the study design that makes the required sample size so much smaller than that in Kramer's study?

(self) paired nature of the observations; also, less stringent beta.

c What do you consider would be a clinically significant advance in sleep onset time?

Whatever. it's your judgement. For me, I would have to look at the tradeoffs.

d "In all 8 subjects sleep onset time was earlier during melatonin treatment than during placebo" [Abstract]
List 3 possible tests of these data, putting them in order of increasing statistical power [do not carry out the tests, but give references]

sign test < signed rank test < paired t-test

e Set up the calculation from which the p<0.01 for the 3.49 versus 2.12 [Table II, sleep onset time, melatonin versus placebo] was derived

d = difference for one subject;
dbar = ave of 8 differences = 3.49 - 2.12;
s = sd of the 8 differences

$t = \frac{dbar}{s/\sqrt{8}}$ *; refer to t table with 7 df (two sided)*

the fact that each half of the difference was based on an average of 4 weeks of sleep logs doesn't get used explicitly, but does get used implicitly in that an average value based on 4 weeks is a lot more stable than a value based on 1 or 2 nights.

10 [10 points] Statistical Power and Sample Size

Suppose that on the basis of observing a person on 10 randomly chosen occasions, you classify the person into one of two types

One who, in the 10 observations,

'+' wore the seat belt 'significantly more often than 50%' i.e. p <0.05 1 sided.
(you infer that the person wears seat belts in a MAJORITY of ALL occasions, not just the 10 observed here)

'–' did not
(because your 'test' is one sided, this category includes the person for whom you might infer that (s)he wears the seat on a 'minority' of ALL occasions and the person for whom you do not have 'sufficient evidence' that (s)he is a 'majority' user)

- On at least how many occasions out of 10 must you observe that the person used a seat belt in order to classify the person as a '+' ? Why?

  Want the smallest significant 'y' i.e. the y such that ...

  Prob (y or more| π =0.5) < 0.05 but Prob (y-1or more | π =0.5) > 0.05

  By trial and error, working down from y=10, to y=9, etc... we find from Table on page 24 of coursepack (or binomial table in text) that with n=10

  Prob (9 or more | π =0.5) = 0.001 + 0.010 = 0.011

  Prob (8 or more | π =0.5) = 0.001 + 0.010 + 0.044 = 0.055

  so we need 9 or more.

- Suppose there are really only 5 groups of persons: those who wear their car seat belt on 0% of all possible occasions, those who do so on 25% of all occasions, those who do so on 50%, on 75%, and those who do so on 100%. If you use the sampling and classification scheme above, what proportion of these different groups of persons will you classify as '+' ? Draw these proportions as a type of 'power curve' below. Make sure to label the axes. If you have time, fill in the values for 55%, 60%, ... 95%.

  again, from Table on page 24 of coursepack (or binomial table in text) that with n=10

  Prob (9 or more | π =0.75) = 0.056 + 0.188 = 0.244

- Repeat the calculations for a system based on samples of size 20

  we need 15 or more out of 20 .

  Check: probabilities of

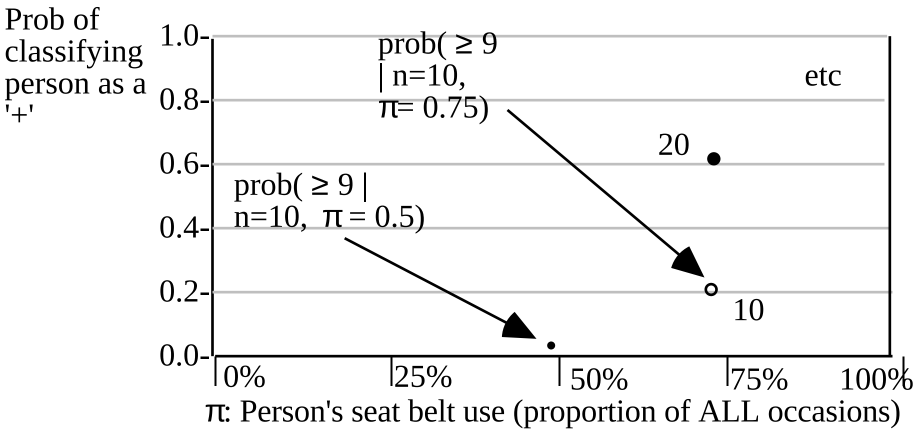
20	*19*	*18*	*17*	*16*	*15*	*14 are..*
0.000	*0.000*	*0.000*	*0.001*	*0.005*	*0.015*	*0.037*

so cumulation (tail) from 15 to 20 is less than 0.05 but from 14 to 20 exceeds 0.05

Fom Table on page 24 of coursepack (or binomial table in text) with n=20

Prob (15 or more | π =0.75) =

0.003 + 0.021 + 0.067 + 0.134 + 0.190 + 0.202 = 0.617


Prob of classifying person as a '+'
1.0-
0.8-
0.6-
0.4-
0.2-
0.0-
prob(≥ 9 | n=10, π= 0.75)
etc
20
prob(≥ 9 | n=10, π = 0.5)
10
0%
25%
50%
75%
100%
π: Person's seat belt use (proportion of ALL occasions)