1. Questions re van Belle et al.

i. For each of Problems 5.1 to 5.11 and 5.13 to 5.16 (p141 to 149) say whether the inference involves a single μ_Y, a single μ_D where D is a *paired* difference $Y_1 - Y_0$, or the difference $\mu_2 - \mu_1$.

ii. Explain why, in the worked examples 5.6 and 5.7, page 136-137, the sample size formula used does not seem to involve σ. Is it a typo, or something else? Hint: look up the term "effect size", used a lot in psychometrics, clinical epidemiology, etc, when using instruments with an arbitrary scale (e.g. GRE scores)

iii. In Comment 2. at the bottom of page 137, the authors suggest a simple rule of thumb to increase the sample sizes to compensate for using $z_{\alpha/2}$ and z_β rather than $t_{\alpha/2}$ and t_β. Sample size tables based on the t distribution, taken from the CRC tables, are given in the attached excerpts from JH's Notes. Does the rule of thumb match what is in these tables? [extensive comparisons not required]

2. What if the primary end-point was the Length of Stay [LOS]?

The following excerpts are from the article Perioperative Normothermia to reduce the incidence of surgical-wound infection and shorten hospitalization. A Kurz et al. N Engl J Med 1996;334:1209-15.

Abstract

Background: Mild perioperative hypothermia, which is common during major surgery, may promote surgical-wound infection by triggering thermoregulatory vasoconstriction, which decreases subcutaneous oxygen tension. Reduced levels of oxygen in tissue impair oxidative killing by neutrophils and decrease the strength of the healing wound by reducing the deposition of collagen. Hypothermia also directly impairs immune function. We tested the hypothesis that hypothermia both increases susceptibility to surgical-wound infection and lengthens hospitalization.

Methods: Two hundred patients undergoing colorectal surgery were randomly assigned to routine intraoperative thermal care (the hypothermia group) or additional warming (the normothermia group). The patients anesthetic care was standardized, and they were all given cefamandole and metronidazole. In a double-blind protocol, their wounds were evaluated daily until discharge from the hospital and in the clinic after two weeks; wounds containing culture-positive pus were considered infected. The patients' surgeons remained unaware of the patients' group assignments.

Results: The mean (SD[1]) final intraoperative core temperature was 34.7(0.6)C in the hypothermia group and 36.6(0.5)C in the normothermia group (P ...). Surgical-wound infections were found in x of 96 patients assigned to hypothermia (a percent) but in only y of 104 patients assigned to normothermia (b percent, P ...). The sutures were removed one day later in the patients assigned to hypothermia than in those assigned to normothermia (P ...), and the duration of hospitalization was ******ed by x.x days (approximately pp percent) in the hypothermia group (P ...).

Conclusions: ....

Methods (2 paragraphs from the Methods section in full text)

The number of patients required for this trial was estimated on the basis of a preliminary study in which 80 patients undergoing elective colon surgery were randomly assigned to hypothermia (mean [SD] temperature, 34.4(0.4)C or normothermia (involving warming with forced air and fluid to a mean temperature of 37(0.3). The number of wound infections (as defined by the presence of pus and a positive culture) was evaluated by an observer unaware of the patients' temperatures and group assignments. Nine infections occurred in the 38 patients assigned to hypothermia, but there were only four in the 42 patients assigned to normothermia (P = 0.16).

Using the observed difference in the incidence of infection, we determined that an enrollment of 400 patients would provide a 90 percent chance of identifying a difference with an alpha value of 0.01. We therefore planned to study a maximum of 400 patients, with the results to be evaluated after 200 and 300 patients had been studied. The prospective criterion for ending the study early was a difference in the incidence of surgical-wound infection between the two groups with a P value of less than 0.01. To compensate for the two initial analyses, a P value of 0.03 would be required when the study of 400 patients was completed. The combined risk of a type I error was thus less than 5 percent.

Comment [JH]: As you can see, they planned the sample size on the basis of their primary endpoint, the incidence of infection. JH does not like the

[1]The NEJM continues to use ±SD. JH has deleted the ± and given the SD for what it is, a positive quabtity!

way they chose the "delta" i.e. as the observed difference in the preliminary study; he prefers to define delta as "the difference that would make a difference – a clinical judgment that also takes *costs and other practical issues* into consideration.

Question: Suppose that hospital administrators consider that shortening of the length of stay [LOS] by 1 day would be quite substantial *if it could be achieved*. Suppose further that in similar admissions last year, the mean(SD) hospitalization was 15(7) days. Calculate the required sample size using the same "90 percent chance of identifying a difference with an alpha value of 0.01" (ignore for now the compensation for the interim analyses).

3. Permutation test rather than paired t-test

See Fisher's application of his permutation test to Darwin's (paired) data on growth of plants. He admits that the "arithmetical procedure of such an examination is tedious." Since the task could get even more tedious if there were greater numbers of pairs involved, an alternative is to *sample* from this permutation distribution.

Use `R` (or `SAS` or `SPSS`) to take a sample of the 2^{15} permutations, and thus of the possible sums, and *estimate* the P-value by calculating what % of them are exceeded by the observed sum.

4. Births after The Great Blackout of 1966

On November 9, 1965, the electric power went out in New York City, and it stayed out for a day – The Great Blackout. Nine months later, newspapers suggested that New York was experiencing a baby boom. The table shows the number of babies born every day during a twenty-five day period, centered nine months and ten days after The Great Blackout.

Number of births in New York, Monday August 1-Thursday August 25, 1966.

Mon	Tue	Wed	Thu	Fri	Sat	Sun
451	468	429	448	466	377	344
448	438	455	468	462	405	377
451	497	458	429	434	410	351
467	508	432	426			

These numbers average 436. This turns out to be not unusually high for New York. But there is an interesting twist: the 3 Sundays only average 357.

i. In a previous assignment, you were asked how likely is it that the average of three days chosen at random from the table will be 357 or less. Most of you set this up as a 1-sample hypothesis test, with

$$H_0: \ \mu_{Sundays} = \mu_{other\ days}, \text{with } \mu_{other\ days} \text{ known to be 436,}$$

$$H_{alt}: \ \mu_{Sundays} < \mu_{other\ days}, \text{with } \mu_{other\ days} \text{ known to be 436.}$$

and your Sunday data consisted of $n = 3$ observations, with $\bar{y}_3 = 357$. You would use the t or z distribution, depending on whether you knew or estimated σ.

Exercise 1: Repeat the testing, but using a *permutation* approach, i.e. enumerate all of the possible random samples of sizes 3 and 22, and determine the fraction of such instances in which $\bar{y}_3 < \bar{y}_{22}$

Exercise 2: Repeat the testing, but using a *permutation* of the *ranks* approach, i.e. enumerate all of the possible random samples of sizes 3 and 22, and determine the fraction of such instances in which $\overline{rank}_3 < \overline{rank}_{22}$, i.e., in which the average rank in the sample of 3 was lower than the average rank in the remaining sample of 22.

In their text *Statistics*, Freedman et al. tell us that "Apparently, the New York Times sent a reporter around to a few hospitals on Monday August 8, and Tuesday August 9, nine months after the blackout. The hospitals reported that their obstetric wards were busier than usual – apparently because of the general pattern that weekends are slow, Mondays and Tuesdays are busy. These "findings" were published in a front-page article on Wednesday, August 10, 1966, under the headline "Births Up 9 Months After the Blackout." This seems to be the origin of the baby-boom myth."

5. Reducing the rate of drop-out from exercise classes

Drop-out from exercise classes is substantial. In a study about which JH was consulted, 4 of 8 exercise classes at U. de M. were randomly assigned to receive weekly counselling by a sports psychologist on how to "hang in there" while the other 4 served as a comparison. The mean number of sessions attended was calculated for each class (the mean for a class would be 20 if all 25 students in the class attended all 20 sessions). The means for the 4 experimental classes were 11.1, 12.2, 9.4, and 11.7; the means for the comparison classes were 9.6, 9.2, 10.3, and 9.7.

i. Carry out a 2-sample t-test.

ii. The PI was very reluctant to use a t-test, since she thought the sample sizes (4 and 4) were too small and she was unable to check (or speculate) that the values come from 2 Normal distributions. Carry out a permutation test instead – assume that the 4 experimental classes were randomly chosen from the 8.

6. The effect of working serial night shifts on the cognitive functioning of emergency physicians

i. The mean day-shift KAIT score was 119.1 (SD=7.7), and the mean night-shift KAIT score was 107.2 (SD=10.2). This difference was significant (mean difference=11.9; 95% confidence interval 7.0 to 16.8; P < 0.001), with the dayshift scores being statistically higher than the night-shift scores" (Abstract; but see also more complete summaries in Table 1)

(a) Reconstruct the 95% CI 7.0 to 16.8 from the summaries given.

(b) State the null and alternative hypotheses tested and verify that "P < 0.001"

(c) Why, in the last row of the Table, doesn't $(7.7^2 2 + 10.2^2)^{1/2}$ equal 9.2 ?

ii. Residents in group B, who were tested first after working night shifts, had a larger difference between their 2 scores than residents in group A, who were tested first on the day shift (night first: mean difference=17.1 [SD=8.6]; day first: mean difference= 6.6 [SD=6.7]; P=.017). On the basis of these scores, the order of testing with the KAIT (night first or day first) did make a difference" [Bottom of page 153 and top of page 154]

(a) Reconstruct the P-value (0.017) from the summaries given.

(b) Explain in words – to a resident who is working the day shift – what the P-value of 0.017 is (after the night shift, don't even try!).

(c) Why did the order of testing make a difference? What is the lesson for investigators who are attracted to the crossover design?

7. Paracetamol and Fever

i. Entry was limited to children with temperatures between 38C and 41C.

Given the mean of 38.9C and the SD of 0.9, what can you say about the shape of the frequency distribution over the 38C-41C interval? (give a sketch)

ii. "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])

iii. "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?

iv. "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"

v. "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.

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

vii. How was the CI calculated?

viii. 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]?

ix. "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?

x. "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?

8. Detectable Differences with available sample size in "Probit II" , ie., followup to "**Promotion of Breastfeeding Intervention Trial (PROBIT)**: A Randomized Trial in Republic of Belarus" [2] [*courtesy M Kramer, R Platt*]

Context: Current evidence that breastfeeding is beneficial for infant and child health is based exclusively on observational studies. Potential sources of bias in such studies have led to doubts about the magnitude of these health benefits in industrialized countries. **Objective:** To assess the effects of breastfeeding promotion on breastfeeding duration and exclusivity and gastrointestinal and respiratory infection and atopic eczema among infants. **Design:** The Promotion of Breastfeeding Intervention Trial (PROBIT), a cluster-randomized trial conducted June 1996December 1997 with a 1-year follow-up. **Setting:** Thirty-one maternity hospitals and polyclinics in the Republic of Belarus. **Participants:** A total of 17 046 mother-infant pairs consisting of full-term singleton infants weighing at least 2500 g and their healthy mothers who intended to breastfeed, 16491 (96.7%) of which completed the entire 12 months of follow-up. **Interventions:** Sites were randomly assigned to receive an experimental intervention (n = 16) modeled on the Baby-Friendly Hospital Initiative of the World Health Organization and United Nations Children's Fund, which emphasizes health care worker assistance with initiating and maintaining breastfeeding and lactation and postnatal breastfeeding support, or a control intervention (n = 15) of continuing usual infant feeding practices and policies. **Main Outcome Measures:** Duration of any breastfeeding, prevalence of predominant and exclusive breastfeeding at 3 and 6 months of life and occurrence of 1 or more episodes of gastrointestinal tract infection, 2 or more episodes of respiratory tract infection, and atopic eczema during the first 12 months of life, compared between the intervention and control groups. **Results:** Infants from the intervention sites were significantly more likely than control infants to be breastfed to any degree at 12 months (19.7% vs 11.4%; adjusted odds ratio [OR], 0.47; 95 confidence interval [CI], 0.32-0.69), were more likely to be exclusively breastfed at 3 months (43.3% vs 6.4%; P¡.001) and at 6 months (7.9% vs 0.6%; P = .01), and had a significant reduction in the risk of 1 or more gastrointestinal tract infections (9.1% vs 13.2%; adjusted OR, 0.60; 95% CI, 0.40-0.91) and of atopic eczema (3.3% vs 6.3%; adjusted OR, 0.54; 95% CI, 0.31-0.95), but no significant reduction in respiratory tract infection (intervention group, 39.2%; control group, 39.4%; adjusted OR, 0.87; 95% CI, 0.59-1.28). **Conclusions:** Our experimental intervention increased the duration and degree (exclusivity) of breastfeeding and decreased the risk of gastrointestinal tract infection and atopic eczema in the first year of life. These results provide a solid scientific underpinning for future interventions to promote breastfeeding.

[2]Kramer Shapiro Collet Ducruet, ... et al. ; [JAMA. 2001;285:413-420.

The **principal objective** of **PROBIT II** is to examine whether the experimental breastfeeding promotion intervention introduced in Belarus in 1996 and 1997 has effects detectable at 6 years of age on atopic disease, cognitive development, behaviour, growth, obesity, and blood pressure. The comparison of the experimental and control groups, when analyzed by intention to treat, will allow the most rigorous examination to date of the causal relationship between prolonged, exclusive breastfeeding and these important health outcomes.

Statistical Aspects: Based on the results of our pilot study random sample of PROBIT participants, we expect 86%, or 14,140, of the 16,442 subjects who completed the 12-month follow-up in Phase I will participate in Phase II. The Table shows the differences in the principal study outcomes detectable with 80% power, based on this sample size and an intention-to-treat analysis, with two different assumptions for the value of the intra-cluster, among-individual correlation (the intraclass correlation coefficient, or ICC): .01 and .03. (These values reflect the range in ICCs from outcomes in Phase I of PROBIT.) As can be seen, the projected sample size is ample for detecting clinically important differences in the principal continuous study outcomes and should detect moderate differences in the proportion of children with wheezing symptoms (based on the ISAAC) questionnaire and positive skin-prick tests.

Table. Control Group Means and SDs, Proportions, and Differences (Δ)Detectable at P = .05 with 80% Power.

Continuous Outcomes	Mean	SD	Δ*	Δ**
IQ (Total/Verbal/Performance)	100	15	1.6	2.6
SDQ	8.6	5.7	0.6	1.0
Systolic blood pressure (mm Hg)	95	10	1.1	1.8
Diastolic blood pressure (mm Hg)	58	12	1.3	2.1
Height (cm)	45	2.2	0.2	0.4
Body mass index (kg/m^2)	15.2	1.8	0.2	0.3
Dichotomous Outcomes	Proportion (%)		Δ*	Δ**
$\geq$ 1 Positive skin-prick test	20%		4%	7%
Wheezing in past 12 months	8%		2.7%	4.1%

Based on intraclass correlation coefficients (ICC's) of 0.01* and 0.03**.

Exercise. Assume a comparison of outcomes in 15 × 500 = 7500 children in 15 hospitals and polyclinics randomly assigned to receive the experimental intervention with those in 15 × 500 = 7500 children in the 15 assigned to receive the control intervention. Calculate the detectable difference for the IQ and wheezing variables.[3]

[3]cf formula in the notes; because of slightly smaller numbers used in this exercise than in grant application, your detectable diffrences will be slightly different.

9. Another Cluster Randomized Controlled Trial

Informing Resource-Poor Populations and the Delivery of Entitled Health and Social Services in Rural India: A Cluster Randomized Controlled Trial.[4]

Context A lack of awareness about entitled health and social services may contribute to poor delivery of such services in developing countries, especially among individuals of low socioeconomic status.
Objective To determine the impact of informing resource-poor rural populations about entitled services.
Design, Setting, and Participants Community-based, cluster randomized controlled trial conducted from May 2004 to May 2005 in 105 randomly selected village clusters in Uttar Pradesh state in India. Households (548 intervention and 497 control) were selected by a systematic sampling design, including both low-caste and midto high-caste households.
Intervention Four to 6 public meetings were held in each intervention village cluster to disseminate information on entitled health services, entitled education services, and village governance requirements. No intervention took place in control village clusters. Main Outcome Measures Visits by nurse midwife; prenatal examinations, tetanus vaccinations, and prenatal supplements received by pregnant women; vaccinations received by infants; excess school fees charged; occurrence of village council meetings; and development work in villages.
Results At baseline, there were no significant differences in self-reported delivery of health and social services. After 1 year, intervention villagers reported better delivery of several services compared with control villagers: in a multivariate analysis, 30% more prenatal examinations (95% confidence interval [CI], 17%-43%; P $<$ .001), 27% more tetanus vaccinations (95% CI, 12%-41%; P(95% CI, 8%-39%; P=.003), 25% more infant vaccinations (95% CI, 8%-42%; P=.004), and decreased excess school fees of 8 rupees (95% CI, 4-13 rupees; P $<$ .001). In a difference- in-differences analysis, 21% more village council meetings were reported (95% CI, 5%-36%; P=.01). There were no improvements in visits by a nurse midwife or in development work in the villages. Both low-caste and mid- to high-caste intervention households reported significant improvements in service delivery.
Conclusions Informing resource-poor rural populations in India about entitled services enhanced the delivery of health and social services among both low- and midto high-caste households. Interventions that emphasize educating resource-poor populations about entitled services may improve the delivery of such services.

Trial Registration clinicaltrials.gov Identifier: NCT00421291

[4]Pandey et al. JAMA. 2007;298(16):1867-1875

Methods: Setting and Sample Selection (from full text)

Our cluster-randomized trial sample size calculations were based on a 5% significance level and 80% power. The sample size and power calculations are driven by the number of village clusters, rather than the number of households per village cluster. For proportional outcomes, to detect a 0.2 increase over a control proportion of 0.5 with 10 households per cluster and a conservative coefficient of variation of 0.5, we estimated needing 94 total clusters (47 per arm). Increasing the number of households above 10 does not significantly decrease the number of village clusters required. For school fees, to detect a 10-rupee decline from a control of 35 rupees with 10 children per cluster, standard deviation of 15 rupees, and a coefficient of variation of 0.5, we estimated needing 82 total clusters. Our actual sample size included 105 total clusters.[5]

Excess school fees were defined as the school fees paid by students minus the legal amount they can be charged (US $1=45 rupees). The unit of analysis for this outcome was individual children. The unit of analysis for other outcomes (eg, visits by nurse midwife, development work in village) was households. For each outcome, we compared intervention and control groups, adjusting standard errors for clustering at the village level. We used the `regress` and `cluster` commands from `Stata` 9.2 statistical software (StataCorp, College Station, Texas) for these analyses. P .05 was set as the threshold for significance. For 5 of 8 outcomes, comparing within-household changes from baseline to follow-up was not possible, because households that reported those outcomes at baseline were often not reporting on the same outcomes at 1 year. For example, a household reporting on prenatal outcomes at baseline would no longer have a pregnant woman to report prenatal outcomes on at 1 year. For these, we additionally conducted a multivariate regression comparing intervention to control at 1 year, using a random- effects model in which random effects are at the village cluster level and standard errors are clustered at the village cluster level. The regression adjusts for total population of the village cluster, district size, household caste, and highest education attained in the household. For the 3 remaining outcomes of visits by nurse midwife, village council meetings, and development work in village, we conducted a within-household difference-indifferences analysis, using a random effects model at the village cluster level and clustering for standard errors at the village cluster level. Focus groups were analyzed by proportion of respondents to questions. Quotations representing dominant themes were recorded.

Exercise. Try to replicate the sample size calculations.

[5]Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. Int J Epidemiol. 1999; 28(2):319-326.