1 True or false, and explain briefly.

a If you add 7 to each value on a list, you add 7 to SD.
FALSE: shifting data does not change variability

b If you double each value on a list, you double the SD
TRUE: doubling the scale doubles the variability

c If you change the sign of each value on a list, you change the sign of the SD
FALSE: Standard Deviation is non-negative by defn.; same whether data "left to right" or "right to left"

d If you duplicate each value in a list, you leave the SD approximately unchanged
TRUE: variability remains the same

e If all the values in a list are positive, you cannot have a SD which is larger than the mean
FALSE: if long R tail. e.g. if 90% of observations are 0.1 and 10% are 1.1, mean is 0.2 but SD is 0.3 (same for obsns. on ≥ 2-point scale)

f Half the values on a list are always below the mean
FALSE: cf. news story on 'below average' children

g In a large list, the distribution of measurements follows the normal curve quite closely
FALSE: "large n does not make a Gaussian distribution". Observations have a pattern of variation of their own,; doesn't change shape if humans observe; behaviour of mean of a large n of indep. observations another story (CLT).

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.
FALSE: it all depends on the pattern of variation; for example, could have 50% at 40 and 50% at 60. See "5 distributions with same mean & SD" in Ch..1.

i The variance of the sum of two random variables is always bigger than the variance of each one.
TRUE if r.v.'s are independent; if sufficiently negatively correlated, it could be less (e.g. amounts of housework done by each of male and female partners)

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.
TRUE: mean increases but 25th, 50th and 75th %-ile unchanged.

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.
FALSE: cannot always draw inferences about "longitudinal" behaviour from cross-sectional data.

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]

MAIN CALCULATION ERROR: principal mixed scales: used 'count' or (0,20) scale for numerator of test statistic, but calculated SD for variation on the 'proportion' (0,1) scale for denominator of test statistic.
LOGICAL ERROR IN INFERENCE: the classic "prosecutor's fallacy" of equating the probability of the data given a hypothesis, with the probability of a hypothesis given the data. May be number of other explanations, such as having had the material explained badly in class, or, as one of you said, "SOMETIMES IT IS EASIER TO BE WRONG BY REASONING THAN BY CHANCE". Or, to quote from p 113 of Ch 9 "Elements of Data Analysis and Inference" in Miettinen's text "Theoretical Epidemiology", ...

> *The P-value is only a statistic, a partial summary of the evidence for or against the denial of the hypothesis in a particular body of data... The probability that the "null hypothesis" is correct is a quantitative expression of the extent to which someone believes in the denial proposition. This depends only partly on the evidence in the data -- expressed not as the P-value but as the likelihood ratio function. An additional determinant is the person's view of the hypothesis apart from the data.*

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 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?

	POPULATION		% dying	# dying		at average age
age group	**LEFT**	**RIGHT**	in next	**LEFT**	**RIGHT**	at death of
	(K)	(K)	5 years	(K)	(K)	
0 to 25	150	1000	0.5%	1.5	5.0	10 years
25 to 50	100	1400	1.5%	1.5	21.0	40 years
50 to 75	32	720	12.5%	4.0	90.0	65 years
75 to 100	2	100	50.0%	1.0	50.0	80 years
number of deaths				8K	166K	
average age at death (years)				**52**	**65**	

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.

Not obvious that he had a comparison group (except in his head)

No mention of: on whom? when? how often? how they would have done without bleeding or with just tlc.

He seems to 'capitalize on chance' by reporting his best day. Length of f-u unclear.

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?

Total vertical progress $VP = P_{D1} + VP_{N1} + ... + VP_{D16} + VP_{N16}$.

$E(VP) = (22 - 12) + ... + (22 - 12) = 160$

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

Remember that it is variances that add, not SD"s

Because VP should be Gaussian (see reasons below), could be 95% sure that it was between 120 & 200 cm from base.

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

Assuming VP is N(μ=160, SD=20), then prob(VP ≥ 150)
= prob {Z ≥ (150–160)/20 } = prob {Z ≥ –0.5 } = 0.69

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.

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

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 37% : statistic or parameter? *STATISTIC, calculated from sample of data.*

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

(ii) sampling variation associated with estimate of % reduction. Basic data on individuals are 1's and 0's [thrombosis or not] . In each study, % reduction is derived from two binomial statistics..

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]
Don't know how symmetric/Gaussian the sampling variation estimate of % reduction would be, but as a first approximation, could expect, even if the reduction were the same in the two subtypes, the random difference in any two samples would be non-zero, and would be Gaussian with SD approximately equal to $\sqrt{11^2 + 13^2} = 17$. *So an observed difference of 49 – 31 = 18 would not be that unusual. One sees same thing if plots the CI's. Moreover, the SD of 17 refers to 2 random samples, not the 2 furthest apart of 3 random samples.*
Some of you took SD of the 3 estimates; but SD associated with each estimate reflects sample sizes etc. Also, SD and SE interchangeable here.

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 components. Calculate SD for
$\frac{1}{3}\text{estimate}_1 + \frac{1}{3}\text{estimate}_2 + \frac{1}{3}\text{estimate}_3$ [§4.3 & 4.64–4.66 should help]

using rules for Var(sum) and var(constant X) and defn. of SD

$var\left[\frac{1}{3}\text{estimate}_1 + \frac{1}{3}\text{estimate}_2 + \frac{1}{3}\text{estimate}_3 \right]$

$= \frac{1}{9}64 + \frac{1}{9}169 + \frac{1}{9}121 = 39.33$ so $SD = \sqrt{39.33} = 6.27$

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 that are 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 This information-weighted average has a lower uncertainty {SD} associated with it than a simple equally-weighted ave.. Calculate SD for $0.52\text{estimate}_1 + 0.20\text{estimate}_2 + 0.28\text{estimate}_3$; compare with SD in d.

$0.52\text{est}_1 + 0.20\text{est}_2 + 0.28\text{est}_3 = 0.52(37\%)+..+0.28(49\%) = \underline{39.16}\%$
$\text{var}[\ 0.52\text{estimate}_1 + 0.20\text{estimate}_2 + 0.28\text{estimate}_3\ ]$
$= 0.52^2\ \text{var}[\text{est}_1] + 0.20^2\ \text{var}[\text{est}_2] + 0.28^2\ \text{var}[\text{est}_3]$
$= 0.2704\ (64) + 0.0400\ (169) + 0.0784\ (121) = \underline{33.55}$
so SD = $\sqrt{33.55} = \underline{5.79}$, *smaller (by definition) than SD above.*

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

f Combine the single estimate for surgical patients and the 42% for medical patients. Calculate its SD. Why does the 'averaging' of the two estimates not diminish the SD very much?
$\text{estimate}_{\text{surg}} = 39.16\%$ with $\text{SD}[\text{estimate}_{\text{surg}}] = 5.79$, var= 33.55;
$\text{estimate}_{\text{med}} = 42.00\%$ with $\text{SD}[\text{estimate}_{\text{med}}] = 17.00$; var=289.00;
optimal weights proportional to $\frac{1}{5.79^2}$ and $\frac{1}{17^2}$ or $\underline{0.90}$ and $\underline{0.10}$
$0.90\text{est}_{\text{surgical}} + 0.10\text{est}_{\text{medical}}$
$= 0.90(39.16\%)+0.10(42\%) \quad = 39.44\%$
Overall estimate closer to 39.16% than 42% because weighted 9:1
$\text{var}[\ 0.90\text{estimate}_{\text{surgical}} + 0.10\text{estimate}_{\text{medical}}\ ]$
$= 0.90^2\ (33.55) + 0.10^2\ (289.00) = 30.07$ so SD = $\sqrt{30.07} = 5.49$

SD only slightly smaller: estimate dominated by estimate$_{surgical}$.

6 A health department serves 50,000 households. As part of a survey, a srs of 400 of these households are surveyed. The average(SD) number of adults in the sample households is 2.35(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]

no. of adults x	% of households with x adults	
1	25	
2	5	mean 2.35
3	25	s.d. 1.11
4	10	
5	5	
	total 100	

We got a variety of shapes of distributions, all with a long right tail.

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

since n = 400 is large, <u>even if x 's are not Gaussian, can invoke CLT</u> and so use large-sample CI for μ_x of form mean $\pm\ z\frac{sd}{\sqrt{n}}$;

$2.35 \pm 1.96\frac{1.11}{\sqrt{400}}$ i.e. 2.35 ± 0.11 or <u>2.24 to 2.46 adults ;</u>

<u>CI for TOTAL No. of Adults</u> : 50 000 times CI for μ_x

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.

Large-sample CI for μ_{tv} of form estimate ± z SE(estimate) ; however, since the 940 observations are not likely to be independent, cannot say that n=940. If there is positive correlation between members in same household, 'real' SE is larger than SD/√940. Put another way, the 'effective sample size' is less than 940 and probably greater than 400. Assuming that that the correlation is positive, we could conservatively calculate the SE with n=400 to get a CI of

$$4.2 \pm 1.96 \frac{2.1}{\sqrt{400}}$$

i.e. 4.2 ± 0.21 or 3.99 to 4.41 hours

This is an example of a cluster sample. Books on survey sampling deal with SE's for estimates derived from such samples.

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 81,411 inches (0% CI).

There is approximately a 100–68 = 32 % chance that we are over-estimating or under-estimating the true elevation by more than 6 inches [± 1 SEM].

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

FALSE; a CI is for a parameter, which in this e.g. is the height of the mountain.

Presumably, arithmetic to average the 25 measurements was done correctly, so should be 100% confident in xbar (for what it is, an ESTIMATE of the parameter)

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

CORRECT; that's what a CI is; [not fussing about 1.96 versus $2\frac{30}{\sqrt{25}}$]

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

FALSE; SD of individual readings was 30 inches; statement mixes up SE [of mean] and SD [of individuals]

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

au contraire!

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".

$H_0: \mu_{\text{regular exercise}} = \mu_{\text{no regular exercise}}$ $H_{alt}: \mu \neq \mu$

$$\text{test statistic} = \frac{132 - 120}{16\sqrt{\frac{1}{32} + \frac{1}{32}}} = 3.0 ,$$

so (whether we use t with 62 df (or M&M's conservative df of 31) or the z distribution, the difference is much more extreme than one would expect under H_0.

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

The fact that the numerical difference is far greater than we would expect if chance (random variability) were the only factor operating does not mean that we can attribute it (or part of it) to the hypothesis that exercise does indeed help to maintain mental ability. This is not an experimental study and those in the exercise group might have chosen to exercise for any of a number of reasons related to what is measured with an IQ test. Maybe those with higher IQ are more inclined to exercise. Just because a study found that the writing skills of Macintosh users are poorer than those of DOS based computers, doesn't mean these Mac users would improve their skills if they switched to DOS machines!

Statistical tests are about the magnitudes of numerical differences, but not about the reasons for them.

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

• design (i) 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 thus will require a smaller sample size than (ii). fewer children overall to test. Also fewer informed consents to obtain.

• design (ii) is easier to carry out (would need to go to more schools for (i); also (ii) allows direct matching on [ie elimination of] age, whereas (i) will require synthetic matching [standardization of tests by age]. As we will see below, it is a little more 'unnatural' to anticipate the 'per unit' variance we should use in sample size calculations for design (i) but that is hardly a good reason to choose design (ii). A few of you mentioned doing this in only a few classrooms but I would be wary of getting effects of better and worse teaching mixed in with it so I wouldn't concentrate the sample too much. I would match roughly on age and I would discuss whether to match on gender. One issue here is whether we would take an only child?

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

(i) paired t-test (ii) a 2-sample t-test for independent samples.

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.

If (i) use formula in page 3 of §7.1 of material; if (ii) use page 2 of §7.2.

Have been given $\Delta=3$;

Take alpha=0.05 two sided [tell her that not all referees and editors will take 1-sided tests].
Take beta=0.2 [what others often use; not a good reason for the choice, but gives a 4/5 chance that study will produce a statistically significant difference if Δ is indeed 3].

If being exact about it, would use t values rather than z-values in sample size calculations but not possible in 1 pass as t value depends on sample size! so use z instead as a first iteration.

The one remaining piece is the 'per unit' variation σ .

For design (ii) we are given $\sigma=10$ so $n= 16(\sigma/\Delta)^2 = 178$ / gp.

For design (i) need some idea of σ_d, the variation across pairs with respect to 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. I expect that estimates of ρ can be found in literature; if not I would put it conservatively at 0.4 or 0.5 [it's probably higher].

For (i) if say take $\sigma_d=11$ so $n= 8(\sigma_d/\Delta)^2 = 108$ pairs.

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?

Margin of error for proportion $= 1.96\sqrt{\frac{p\,[1-p]}{n}}$

p = 0.632, 0.583 & 0.655 ; n = 590, 636 & 629 .

so 0.039, 0.038 and 0.037 (or 3.9%, 3.8% and 3.7%) respectively.

CI for Δ *:* $0.632-0.655 \pm z\sqrt{SE^2 + SE^2}$

$-2.3\% \pm 5.4\%$ *i.e. RM could be 7.7% to +3.1%*

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.

If p same, and n reduced to n'=0.81n, then $\sqrt{n'} = 0.9\sqrt{n}$*, so that SE and margin of error are increased by a factor of 1/0.9 or approximately 11%..The new margins of error become 1.11 times the margins with n.*

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?

Suppose that [if one studied so many patients that sampling variation was eliminated] there was a non-zero difference of a certain magnitude in treatment outcome. The chance to show a statistically significant difference between results in 2 samples, when this [unknowable] magnitude difference prevails, is affected by sampling variation. The more the sampling variation, and with statistical tests giving 'chance alone' the 1st opportunity to explain observed differences, the more difficult it is for investigators to declare that the difference they see is "statistically significant".

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 in total.

Use analogy of diagnostic tests, but for aggregates rather than individuals

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?

Use titration to get to Δ *where no longer indifferent. If lumpectomy better than r.m., then not much question about which to choose at a personal level; but if lumpectomy gives poorer survival but better quality of life, there is a tradeoff. Say r.m. had a 60% survival and lumpectomy a 59% survival, would you take lumpectomy anyway? If it were 60–58?, 60–57?*
.

---

[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?

t test for 2 independent samples; Use formula on page __ of §7.2
Even simpler: n per group = 16 s^2 over d^2 = 16 times 0.4^2 over 0.3^2 = 29

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) = 0.4343ln(cotinine).

YES. They would simply have a mean and SD 1/0.4343 or 2.3 times larger than what they would be on the log 10 scale.

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]?

Reverse SEM = SD / √n to get SD = SEM • √n. So,, SD = 1.5 • √36 = 9.
SD not helpful for individual variation because distribution skewed.

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?

(i) positive linear relation; above(below)-average values of tend to be paired with above(below)-average values of ..
(ii) prob(r this extreme | correlation in "population" is zero)

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"?

See discussion.

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".
IF, in the world at large, there were NO relationship between nicotine in mother and infant, then the chance of getting a correlation as impressive as 0.49 would less than 1 in 1000. SEE POSTSCRIPT BELOW

g "Maternal concentrations of nicotine were invariably higher than neonatal levels (P<0.001)" [next sentence]. Since this certainly isn't the case for all 94 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?
• Paired t-test on the 36 differences between level in mother and level in her infant; this tests whether mean difference is zero.
• Sign Test on these 36 differences. This tests the hypothesis that the median difference is zero [i.e. that the difference is equally likely to be positive or negative]
[Ch 13 of A&B: signed rank test .. nonparametric analog of paired t-test]

h In plain words, what is meant by the phrase "concentrations of cotinine did not differ significantly between mothers and infants"?
Given the sample sizes, and the inter-individual variation, the difference in means of the 2 samples was no bigger that one might expect if in fact there were no difference in population means [*by "no bigger than we might expect", we usually mean "no bigger than we would see in 95% of samples of this size"]*

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?
Yes, t-test for 2 independent samples: $t = \frac{0.60 - 0.26}{\sqrt{SEM^2 + SEM^2}}$

One would have to decide between a common s in the calculation of each SEM or a separate variance t-test with reduced df.
Given the very skewed nature of the cotinine concentrations, and the sample sizes of 23 and 35, some might be worried that the CLT would not be sufficient to guarantee that the t-tables would be accurate. Two options to minimize / eliminate these concerns would be to (i) transform the concentrations to log concentrations [they are certainly more skewed on the regular scale] and (ii) "go non-parametric" as the authors say they did.

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?
1 SEM. Given authors' worry about Gaussian-based tests, wonder why they used SE's [if tests not very accurate, neither are CI's]. Also, if SE's to be useful, au's should plot CI's (with 1.5 SE's or so for each]. Would have simply plotted the raw data as dot-diagrams on log scale, added the median, and let reader make the inter-ocular traumatic test [mean already in text, but not very helpful in concentration scale].

Postscript on P-values

The one type of answer that frustrated me the most was that given as an explanation of a statistically significant difference, where e**xplanations tended to mix probability statements about hypotheses with what should be (conditional))probability statements about data under, i.e. given, the null hypothesis**. Typical examples as an interpretation of a difference with P<0.001 : " There is less than 1 chance in 1000 that the means (μ's) are different " or "There is only a 0.001 chance that this correlation was observed due to chance" or "you have a probability leas than 1/1000 that there is no correlation between x and y" or ""it indicates a relationship that has only a 1/000 chance of arising from chance and 999/1000 chance of arising from the model (presume person meant an alternative to the null)" or "il y a peu de chance que ce résultat ait été obtenu par l'effet du hasard seulement" or "It is almost impossible that the null hypothesis is true" When the difference was not statistically significant students wrote "there is a considerable chance that the small difference observed between groups is due to random sampling".

An explanation of a P-value should have the word IF in it somewhere. If you don't like saying "IF the null hypothesis were true,..." you can vary it a bit . Try "even if there were no difference in the mean levels in the population at large, the probability of observing --- in two samples of the size we had -- as big a difference as we got (or one bigger) in our samples is less than .... . Lots of writers use "chance alone" as a kind of shorthand, but please do it cleverly so as to main the conditional sense of the statement. For example, "If chance (random variation) were the only factor operating, the probability that we would observe ..." or "the difference was bigger than one would expect if means were really the same and chance was the only factor operating.." or "This difference is beyond the 95% limits of sampling variation (alone)". This last one finesses it a little without actually having to say "IF".

Read the section in Miettinen concerning this reversal of the meaning of the p-value. Lots of otherwise erudite people fall into this trap. Keep a watch out for it! And whenever you have to put a p-value into words, stop andthink of the archbishop of Canterbury and the "prosecutor's fallacy".