Contents

1 Notes on van Belle Chapter

[Other references: Armitage & Berry Ch 7.2; Colton Ch 4]

Main reference: van Belle Chapter 5.2, 5.6 and 5.8

- Although Note 5.2.1 asks you to distinguish two theoretical situations: $\sigma_1 = \sigma_2 = \sigma$, versus $\sigma \neq \sigma_2$,these two are unfortunately seldom clearly distinguishable in practice. The **test of equal variances (section 5.7) is not very accurate**, and can easily be distorted by non-normaility. JH advises doing both the common-variance and separate-variance tests and reporting the less extreme p-value.
- Think of s_p^2, the "**pooled**" estimate of σ^2 [van Belle, row 6 in Table 5.1] as a **weighted average** of s_1^2 and s_2^2]?
- van Belle mentions one adjusted d.f. at the bottom of page 139. Software packages use variants of the **Welch-Satterthwaite[1] approximation**.

  `SAS PROC TTEST` "computes the group comparison t statistic based on the assumption that the variances of the two groups are equal. It also computes an approximate t based on the assumption that the variances are unequal (the Behrens-Fisher problem). The degrees of freedom and probability level are given for each;Satterthwaite's (1946) approximation,

$$df = [((w_1 + w_2)^2)/(([[(w_1^2)/(n_1 - 1)] + [(w_2^2)/(n_2 - 1)]))]$$

  where $w_1 = [(s_1^2)/(n_1 - 1)]$, and $w^2 = [(s_2^2)/(n|2 - 1)]$. is used to compute the degrees of freedom associated with the approximate t. In addition, you can request the Cochran and Cox (1950) approximation of the probability level for the approximate t."

  `R`: if in `t.test`, one sets the logical variable var.equal to TRUE then the pooled variance is used to estimate the variance; otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.

- **Example 5.4 (heights of husband-wife pairs)**: Since his chapter deals with differences of *independent* random variables, he apologizes for being possibly slightly unrealistic when **he *supposes* that husband-wide pairs are *formed independent of stature***.

  Francis Galton[2] studied Marriage Selection and reported that "whatever may be the sexual preferences for similarity or for contrast, I find little indication in the average results obtained from a fairly large number of cases of any single measurable personal peculiarity, whether it be *stature*, temper, eye-colour, or artistic tastes, in influencing marriage selection to a notable degree. Nor is this extraordinary, for though people may fall in love for trifles, marriage is a serious act, usually determined by the concurrence of numerous motives. Therefore we could hardly expect either shortness or, tallness, darkness or lightness in complexion, or any other single quality, to have in the long run a large separate influence."

  Galton found **a correlation of only 0.10** or so between the heights of fathers and mothers, and so **"I am therefore content to ignore it, and to regard the Statures of married folk just as if their choice in marriage had been wholly independent of stature."**. However, we discovered that the correlation of poorly measured variables (such as self-reported heights in Galton's study) is less than it should be.

  Karl Pearson, who had his graduate students carefully measured the heights of over 1000 husband-wife pairs[3], found a correlation of approximately 0.30 and commented that "there is a very sensible resemblance in size between hustand and wife, which *à priori* I should have said was hardly conceivable."

  It is not clear how strong the correlation is in today's societies. But it is a pity that van Belle did not check it out: he did however concede that his supposition (of independence) was "probably contrary to societal *mores*."

[1] Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. Biometrics Bulletin 2: 110-114.
Welch, B. L. 1947. The generalization of student's problem when several different population variances are involved. Biometrika 34: 28-35.

[2] Natural Inheritance, 1889, chapter VII.

[3] Table II, page 373 in Pearson, K., and Lee, A. (1903), On the Laws of Inheritance in Man: I. Inheritance of Physical Characters, Biometrika, 2, 357462.

2 Unequal Sample Sizes ($n_1 \neq n_2$)

2.1 Effect of unequal sample sizes on *precision* of estimated differences

If we write the SE of an estimated difference in mean responses as $\sigma \times (1/n_1 + 1/n_2)^{1/2}$, where σ is the (average) per unit variability of the response, then we can establish the following principles[4]:

1. **If costs and other factors (including unit variability) are equal, and if both types of units are equally scarce or equally plentiful,** then for a given total sample size of $n = n_1 + n_2$, an equal division of n i.e. $n_1 = n_2$ is preferable since it yields a smaller SE(estimated difference in means) than any non-symmetric division. However, the SE is relatively unaffected until the ratio exceeds 70:30. This is seen in the following table which, assuming $\sigma = 1$ (arbitrary), gives the value of SE(difference in means) $= (1/n_1 + 1/n_2)^{1/2}$ for various combinations of n_1 and n_2 adding to 100 (the 100 is also arbitrary).

n_1	n_2	SE (diff. in means) $[(1/n_1 + 1/n_2)^{1/2}]$	% Increase in SE over $SE_{50:50}$
50	50	0.200	—
60	40	0.204	2.1%
65	35	0.210	4.8%
70	30	0.218	9.1%
75	25	0.231	15.5%
80	20	0.250	25.5%
85	15	0.280	40.0%

2. **If one type of unit is much scarcer, and thus the limiting factor,** then it makes sense to choose all (say n_1) of the available scarcer units, and some $n_2 \geq n_1$ of the other type. The greater is n_2, the smaller the SE of the estimated difference. However, there is a 'law of diminishing returns' once n_2 is more than a few multiples of n_1. This is seen in the following table which gives the value of $(1/n_1 + 1/n_2)^{1/2}$ for n_1 fixed (arbitrarily) at 100, and n_2 ranging from $1 \times n_1$ to $100 \times n_1$; again, we assume $\sigma = 1$.

[4]Note: these principles apply to both measurement and count data

n_1	n_2	Ratio (K)	$SE(\hat{\mu_1} - \hat{\mu_2})$	$SE_{K:1}$ as % of $SE_{1:1}$	$SE_{K:1}$ as % of $SE_{\infty:1}$*	$I_K : I_\infty$
50	50	1.0	0.2000	—	1.414	0.50
50	75	1.5	0.1825	91.3%	1.290	0.60
50	100	2.0	0.1732	86.6%	1.225	0.67
50	150	3.0	0.1633	81.6%	1.155	0.75
50	200	**4.0**	0.1581	79.1%	1.118	**0.80**
50	250	5.0	0.1549	77.5%	1.095	0.83
50	300	6.0	0.1527	76.4%	1.080	0.86
50	400	8.0	0.1500	75.0%	1.061	0.89
50	500	10.0	0.1483	74.2%	1.049	0.91
50	1000	20.0	0.1449	72.4%	1.025	0.95
50	5000	100.0	0.1421	71.1%	1.005	0.99
50	∞	∞	0.1414	70.7%	1.000	1.00

This table is the basis for the 'epidemiologic rule of thumb' that a $n_2 : n_1$ ratio of more than 4 is wasteful. The 4 seems to have arisen by focusing on *80% efficiency*: if we use **I = Information, i.e., Inverse of Variance** – as the criterion, one can see that, relative to the perfect (100%) information with an infinite $n_2 : n_1$ ratio, the information with a ratio of K is $K/(K+1)$, which indeed attains a value of 0.8 with $K = 4$.

2.2 *Sample size calculations* when using unequal sample sizes to estimate / test difference in 2 means

For power (sensitivity) $1 - \beta$, and specificity $1 - \alpha$ (2-sided), the sample sizes n_1 and n_2 have to be such that

$$Z_{\alpha/2} \times SE(\bar{y_1} - \bar{y_2}) - Z_\beta \times SE(\bar{y_1} - \bar{y_2}) = \Delta = \mu_2 - \mu_1$$

(if $\beta < 0.5$, then Z_β will be negative). If we assume equal per unit variability, σ, of the y's in the 2 populations, we can write the requirement as

$$Z_{\alpha/2} \times \sigma \times (1/n_1 + 1/n_2)^{1/2} - Z_\beta \times \sigma \times (1/n_1 + 1/n_2)^{1/2} = \Delta$$

If we rewrite $(1/n_1 + 1/n_2)^{1/2}$ as $([1/n_1] \times [1/n_1 + 1/n_2])^{1/2}$ and rearrange the inequality, we get

$$n_1 = \left\{1 + \frac{n_1}{n_2}\right\}(Z_{\alpha/2} - Z_\beta)^2 \left\{\frac{\sigma}{\Delta}\right\}^2.$$

or, denoting n_2/n_1 by K,

$$n_1 = \left\{1 + \frac{1}{K}\right\}(Z_{\alpha/2} - Z_\beta)^2\left\{\frac{\sigma}{\Delta}\right\}^2.$$

i.e., with $n_{smaller}$ denoting the smaller sample size, ...

$$n_{smaller} = \left\{\frac{K+1}{K}\right\}(Z_{\alpha/2} - Z_\beta)^2\left\{\frac{\sigma}{\Delta}\right\}^2.$$

If $K = 1$, so that $n_1 = n_2$, then we get the familiar "2" at the front of the sample size formula for *each* group.

3 Power / Precision / Sample Size: Correlated responses; cluster samples

Suppose that, instead of n independent responses (y's) from population (condition) 1, and a separate n independent responses from population (condition) 2, we have responses on $n = m \times k$ individuals from m clusters of size k each, from population (condition) 1, and on a separate set of n individuals from m clusters of size k each, from population (condition) 2. Examples might be responses of persons in same family or school or medical practice—or even several responses for each subject. Suppose the intra-class correlation is

$$icc = \sigma_b^2/(\sigma_b^2 + \sigma_w^2),$$

where σ_b^2 denotes the variance of the (true) cluster means [within the same population], and σ_w^2 denotes the within-cluster variance. Assume that the *icc* has the same value for each population.

Let $\bar{y}_{1i}$ be the mean of the k y's measured on the ith sampled cluster from population 1 ($i = 1, \ldots, m$), and let $\bar{y}_2$ be the mean of the k values measured for the ith cluster sampled from population 2.

Define $\bar{\bar{y}}_1 = (1/m)\sum_i \bar{y}_{1i}$ for the sample from population 1, and correspondingly $\bar{\bar{y}}_2$ for the one from population 2.

Then $Var[\bar{\bar{y}}_1] = (1/m)^2 \times \sum_1^m Var[\bar{y}_{1i}]$.

Now, $Var[\bar{y}_{1i}] = \sigma_b^2 + (1/k)\sigma_w^2$,

so $Var[\bar{\bar{y}}_1] = (1/m)^2 \times m \times \{\sigma_b^2 + (1/k)\sigma_w^2\} = (1/m)\sigma_b^2 + (1/\{m \times k\})\sigma_w^2$.

Thus

$$Var[\bar{\bar{y}}_1] = \frac{\sigma_b^2}{\text{no. of clusters}} + \frac{\sigma_w^2}{\text{no. of individuals}}.$$

If we had responses from n *unrelated* individuals, i.e., if we had $m = n$ and $k = 1$, then

$$Var[\bar{\bar{y}}_1] = \frac{\sigma_b^2 + \sigma_w^2}{\text{no. of individuals}}.$$

The ratio of the variance with $n = m \times k$ to that with $n = n \times 1$ is therefore

$$\left\{\frac{\sigma_b^2}{m} + \frac{\sigma_w^2}{mk}\right\} \div \left\{\frac{\sigma_b^2 + \sigma_w^2}{mk}\right\} = \frac{k\sigma_b^2 + \sigma_w^2}{\sigma_b^2 + \sigma_w^2} = 1 + (k-1)\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} = 1 + (k-1)icc.$$

i.e., the **Variance (or Sample Size) Inflation Factor** (VIF or SSIF) is

$$\boxed{\textbf{VIF} = \text{SSIV} = 1 + (k-1) \times icc.}$$

Thus, if the value of *icc* is positive, there is less information in a cluster sample of a total of n individuals than there would be in a sample of n unrelated individuals. However, the greater amount of information obtained from a sample of n unrelated individuals might well cost a lot more to obtain, and so the cluster sampling approach may be the more efficient option. In some instances, it may be that the intervention is carried out at the level of the cluster, and it would not make sense to study just one individual per cluster.

A positive correlation doesn't always *increase* variance. It depends on how you use it!

Paul Burton[5] puts it nicely...

> It is clear that if a standard statistical analysis which assumes all observations to be independent is performed on repeated measures (or other correlated) data when the intraclass correlation is positive, results may be misleading. For example, estimated standard errors are likely to be too small, the analysis will effectively assume that there is more information in the data than there really is. Such an analysis has been referred to as naive pooling.
>
> Given that correlation can lead to a loss of information, it may seem surprising that repeated measures designs are used so commonly. **However, when interest centres on *a change in response under different conditions or over time*, the [longitudinal] correlation between repeated observations means that within-person changes can be highly informative because they minimize the "noise" arising from between-person variability**.

[5] *Statistics in Medicine* 17, 1261-1291 (1998) Tutorial in Biostatistics: Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level modelling. Paul Burton, L Gurrin, & P Sly.

> Thus, if one wished to test a new drug purporting to increase height in middle-aged adults, the fact that height is essentially constant in this age group means that the change in height within subjects (before drug versus after) will provide a powerful test of efficacy. In such circumstances, ignoring the correlation structure can waste important information and can make standard errors too large, as when an unpaired t-test is used on paired data with a positive intraclass correlation.

E.S. Pearson, in his appreciation "Student as a Statistician"[6] gives us a good example from Student's writings:

> One of the striking characteristics of these papers, also of course evident in correspondence, was the simplicity of the statistical technique he used. The mean, the standard deviation and the correlation coefficient were his chief tools; hardly adequate for treating specialized problems it might be thought; yet how extremely effective in fact in his skilled hands! **There is one very simple and illuminating theme which will be found to run as a keynote through much of his work, and may be expressed in the two formulae:**
>
> $$\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y + 2\rho\sigma_x\sigma_y$$
>
> $$\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y - 2\rho\sigma_x\sigma_y$$
>
> Perhaps we may count as one of his big achievements the demonstration in many fields of the meaning of the second of these short equations; as he wrote in 1923 [p. 273, "On testing varieties of cereals." Biometrika, Vol 15] but with modified notation:
>
> "**The art of designing all experiments lies even more in arranging matters so that ρ is as large as possible than in reducing σ^2_x and σ^2_y.**
>
> It is a simple idea, certainly, but I cannot doubt that its emphasis and amplification helped to open the way to all the modern developments of analysis of variance, and there may be some who have felt that where this technique runs a risk of defeating its ends by over-elaboration is just where that simple maxim has been set on one side.

[6] Biometrika, Vol. 30, No. 3/4. (Jan., 1939), pp. 210-250.

4 "Eye test" using overlap of 2 indep't CI's

```
|----------x----------|
                                        Overlapping CI's
          |----------x----------|
```

How far apart do two *independent* $\bar{y}$'s, say $\bar{y}_1$ and $\bar{y}_2$ to be for a formal statistical test, using say an $\alpha = 0.05$, two sided, of $\mu_1 = \mu_2$, to be to be statistically significant? **If their associated CI's overlap, does that mean the difference between them is not statistically significant?** [7]

I using a z-test, they will be significantly different if

$$|\bar{y}_1 - \bar{y}_2| \geq 1.96 \times \{(SE[\bar{y}_1])^2 + (SE[\bar{y}_1])^2\}^{1/2}$$

If $SE[\bar{y}_1]$ and $SE[\bar{y}_w]$ are about the same size (as they would be if the 2 n's, and the per-unit variability, were about the same), then, denoting each SEM by $SE[\bar{y}_{each}]$, they are significant if...

$$|\bar{y}_1 - \bar{y}_2| \geq 1.96 \times \{2 \times (SE[\bar{y}_{each}])^2\}^{1/2}.$$

i.e.

$$|\bar{y}_1 - \bar{y}_2| \geq 1.96 \times 2^{1/2} \times SE[\bar{y}_{each}],$$

or...

$$|\bar{y}_1 - \bar{y}_2| \geq 2.77 \times SE[\bar{y}_{each}].$$

If using t rather than z, the multiple would be somewhat higher than 1.96, so that when multiplied by $2^{1/2}$, it would be higher than 2.77, closer to 3. Thus a *rough* answer to the question could be taht they are significantly different if

$$|\bar{y}_1 - \bar{y}_2| \geq 3 \times SE[\bar{y}_{each}].$$

This means that **even when two** $100(1-\alpha)\%$ **CI's overlap slightly**, as above, the difference between the two means could be statistically significant at the α level. This is why Lincoln Moses, in his article on graphical displays (see reserve material), advocates plotting the 2 CI's formed by

$$\bar{y}_1 \pm 1.5 \times SE[\bar{y}_1] \qquad \text{and} \qquad \bar{y}_2 \pm 1.5 \times SE[\bar{y}_2]$$

[7] See Wolfe R, Hanley J. "If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2." Canadian Med Assoc. J. 2002 Jan 8;166(1):65-66.

This way, we can be reasonably sure that if the CI's do not overlap (i.e., if $\bar{y}_1$ and $\bar{y}_2$ are more than 3 $SE[\bar{y}_{each}]$'s apart) then the difference between them is statistically significant at the $\alpha = 0.05$ level, and *vice versa*.

Notes:

- $estimate \pm 1.5SE[estimate]$ corresponds to an 86% CI if using Z distribution.
- The above logic applies for other symmetric CI's too.

5 Permutation tests

From Section 21 "**Test of a Wider Hypothesis**" beginning on page 44 of Chapter III of Fisher's Design of Experiments

> It has been mentioned that "Student's" t test, in conformity with the classicaI theory of errors, is appropriate to the null hypothesis that the two groups of measurements are samples drawn from the same **normally** distributed population. This is the type of null hypothesis which experimenters, rightly in the author's opinion, usually consider it appropriate to test, for reasons not only of practical convenience, but because the unique properties of the normal distribution make it alone suitable for general application.
>
> There has, however, in recent years, been a tendency for theoretical statisticians, not closely in touch with the requirements of experimental data, to stress the element of normality, in the hypothesis tested, as though it were a serious limitation to the test applied. It is, indeed, demonstrable that, as a test of this hypothesis, the exactitude of "Student's" t test is absolute.
>
> It may, nevertheless, be legitimately asked whether we should obtain a materially different result were it possible to test the **wider hypothesis** which merely asserts that the two series are drawn from the **same** population, **without specifying** that this is **normally distributed.**
>
> In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. The arithmetical procedure of such an examination is tedious, and we shall only give the results of its application in order to show the possibility of an independent check on the more expeditious methods in common use.
>
> **On the hypothesis that the two series of seeds are random samples from identical populations, and that their sites have been assigned to members of each pair independently at random, the 15 differences of Table 3 would each have occurred with equal frequency with a positive or with a negative sign.** Their sum, taking account of the two negative signs which have actually occurred, is 314, and we may ask how many of the 2^15 numbers, which may be formed by giving each component alternatively a positive and a negative sign, exceed this value. Since *ex hypothesi* each of these 2^{15} combinations will occur by chance with equal frequency, a knowledge of how many of them are equal to or greater than the value actually observed affords a direct arithmetical test of the significance of this value. *It is easy to see* [JH: typical Fisher phrase!] that if there were no negative signs, or only one, every possible combination would exceed 314, while if the negative signs are 7 or more, every possible combination will fall short of this value. The distribution of the cases, when there are from 2 to 6 negative values, is shown in the following table :-

TABLE 4

Number of combinations of differences, positive or negative, which exceed or fall short of the total observed

Number of negative values.	>314	= 314	<314	Total.
0	1	...	...	1
1	15	...	...	15
2	94	1	10	105
3	263	3	189	455
4	302	11	1,052	1,365
5	138	12	2,853	3,003
6	22	1	4,982	5,005
7 or more	...	...	22,819	22,819
Total	835	28	31,905	32,768

In just 863 cases out of 32,768 the total deviation will have a positive value as great as or greater than that observed. In an equal number of cases it will have as great a negative value. The two groups together constitute 5·267 per cent. of the possibilities available,

a result very nearly equivalent to that obtained using the t test with the hypothesis of a normally distributed population. Slight as it is, indeed, the difference between the tests of these two hypotheses is partly due to the continuity of the t distribution, which effectively counts only half of the 28 cases which give a total of exactly 314, as being as great as or greater than the observed value.

Both tests prove that, in about 5 per cent. of trials, samples from the same batch of seed would show differences just as great, and as regular, as those observed; so that the experimental evidence is scarcely sufficient to stand alone. In conjunction with other experiments, however, showing a consistent advantage of cross-fertilised seed, the experiment has considerable weight ; since only once in 40 trials would a chance deviation have been observed both so large, and in the right direction.

(omitted... a paragraph on a continuity correction)

21.1. "Non-parametric" Tests

In recent years tests using the physical act of randomisation to supply (on the Null Hypothesis) a frequency distribution, have been largely advocated under the name of "Non-parametric" tests. Somewhat extravagant claims have often been made on their behalf. The example of this Chapter, published in 1935, was by many years the first of its class. The reader will realise that it was in no sense put forward to supersede the common and expeditious tests based on the Gaussian theory of errors. The utility of such non-parametric tests consists in their being able to supply confirmation whenever, rightly or, more often, wrongly, it is suspected that the simpler tests have been appreciably injured by departures from normality.

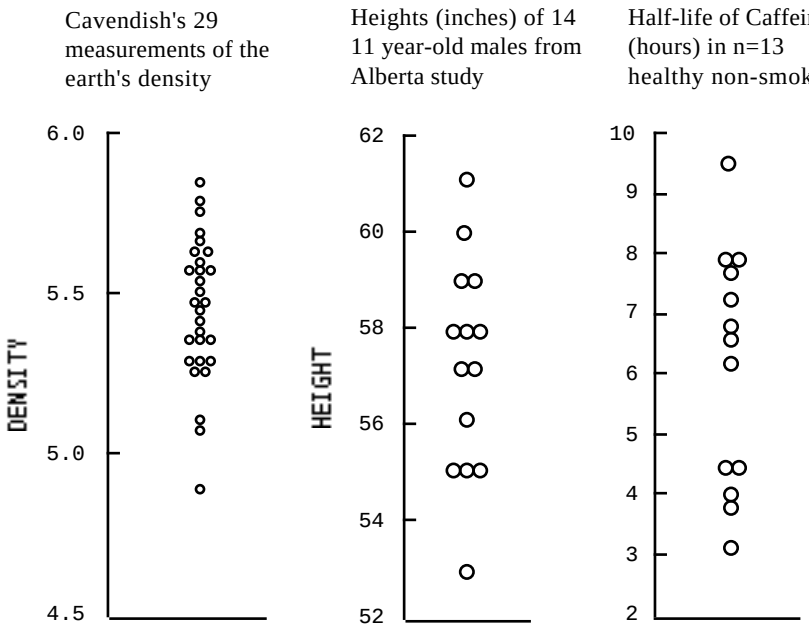
They assume less knowledge, or more ignorance, of the experimental material than do the standard tests, and this has been an attraction to some mathematicians who often discuss experimentation without personal knowledge of the material. In inductive logic, however, an erroneous assumption of ignorance is not innocuous ; it often leads to manifest absurdities. Experimenters should remember that they and their colleagues usually know more about the kind of material they are dealing with than do the authors of text-books written without such personal experience, and that a more complex, or less intelligible, test is not likely to serve their purpose better, in any sense, than those of proved value in their own subject.

.

Note from JH: There is a corresponding permutation test for 2 *independent* samples. Both permutation tests use the ***raw data***, not the *ranks*.

6 t-based CI/test re μ and $\mu_2 - \mu_1$: by regression

6.1 Inference regarding a single μ: data from 1- (or paired-) sample



Statistics

n	... 29	 14	 13
Min	..... 4.88	..... 53.00	..... 9.40
Max	..... 5.85	..... 61.00	..... 5.95
Mean ($\bar{y}$)	..... 5.45	..... 57.21	..... 5.95
Var (s^2)	..... 0.0488	..... 4.9506	..... 3.9460
SD (s)	..... 0.22	..... 2.22	..... 1.99

Least Squares Estimate of μ:

$\sum(y-\bar{y})^2$ is smaller than $\sum(y - \textit{any other central value})^2$.

That's why we can call the statistic $\bar{y}$ the Least Squares estimator of μ. (see applet on best location to wait for elevator in Ch 1 Resources for 607, and 'elevator article' in Ch 1 of Course 697; see also applets in Ch 10 for 607)

Statistical Model:

$$y = \mu + \epsilon$$
$$\epsilon \sim ?(0, \sigma)$$

"Minimum Requirements" for Least Squares Estimation *per se*:

There is *no requirement* that $\epsilon \sim N(0,\sigma)$. Later, for statistical inferences about the parameters being estimated, the inferences may be somewhat inaccurate if n is small and the distribution of the ϵ's is not $N(0,\sigma)$ or if the ϵ's are not independent of each other.

Fitting (i.e. calculating parameter estimates of) model for height:

By calculator (or `SAS PROC MEANS` or `mean` and `var` functions in `R`):

$$\bar{y} = \frac{\sum y}{n} = 57.21; \quad s^2 = \frac{\sum(y-\bar{y})^2}{n-1} = 64.357/13 = 4.95 \quad \rightarrow \quad s = 2.22.$$

From `R`.. `summary( lm(height ~ 1) )`

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.2143 -1.9643  0.2857  1.5357  3.7857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    57.21       0.59   96.22   <2e-16 ***

Residual standard error: 2.22 on 13 degrees of freedom
```

Finding parameter estimates on output of statistical package

If you compare with the calculations above, you will readily identify the estimate $\bar{y} = 57.21$ for the μ parameter.

But what is the estimate of the σ^2 or σ parameter? We know from our calculator that $\hat{\sigma} = 2.22$. In the `R` output (SAS output later!), this estimate is given under the less familiar[8] term Residual *standard error*. You can think of each $(y-\bar{y})$ as the 'residual' variation from the mean $\bar{y}$, and you can therefore call $\sum(y-\bar{y})^2$ the Sum of Squares of the Residuals, or Residual Sum of Squares for short.

[8] Residual *standard deviation*, or *Root Mean Square Error, RMSE,* would confuse less. Systat uses the term Standard Error of Estimate; SPSS uses this 'SEE' terminology too.

What of the other items on the output?

* What is `Std. Error = 0.59465` ? It is the SE of `Intercept` i.e. of $\bar{y}$.

It is what we call the Standard Error of the Mean, or 'SEM' for short, given by the formula

$$\text{Standard Error of Mean} = \text{SEM} = \text{SE}(\bar{y}) = s/n^{1/2} = 2.22/14^{1/2} = 0.59.$$

* What is `t value = 96`?

It is the test statistic corresponding to the test of whether the underlying parameter (μ in our case) is ZERO i.e. of the $H_0 : \mu = 0$. Of course, the computer programmer doesn't know what μ refers to, or that the mean height of 11 year old boys in Alberta is, by definition, greater than zero. Since we might have a case where there was genuine interest in the H_0 that $\mu = 0$ or some other value[9], we will show where $t = 96$ came from: remember from earlier the 1-sample t-test and the formula

$$t = (\bar{y} - 0)/SE(\bar{y}) = 57.21/0.59 = 96.22.$$

* What is `Pr(>|t|) <2e-16` ?

It is the P-value obtained by calculating the probability that an observation from the t distribution with $n - 1 = 13$ df would exceed 96.22 in absolute value.

Fitting "the beginning of all regression models" using `SAS`:

```
proc reg data=sasuser.alberta; model height = ;
```

JH discovered this way of calculating $\bar{y}$ by accident – he forgot to put terms on the right hand side of the `model` statement!

The model is simply

$$y = \mu + \epsilon$$

but it can be thought of as

$$y = \mu \times 1 + \epsilon$$

or

$$y = \mu \times x_0 + \epsilon.$$

where $x_0 \equiv 1$(a constant); it is as though we have set it up so that the "predictor variable" x_0 in the regression equation is always 1. Then μ is the parameter to be estimated.

[9] e.g., We might ask if Cavendish's measurements of the Earth's density are compatible with today's accepted value of 5.518.

Some software programs insist that you specify the constant; others assume it unless told otherwise.

* Output from `SAS PROC REG` Dependent Variable: `height`

Analysis of Variance*

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	0	0.000	.	.	.
Error	13	64.357	4.95		
C Total	13	64.357			

Root MSE	2.22	R-square	0.0000
Dep Mean	57.21	Adj R-sq	0.0000
C.V.	3.89		

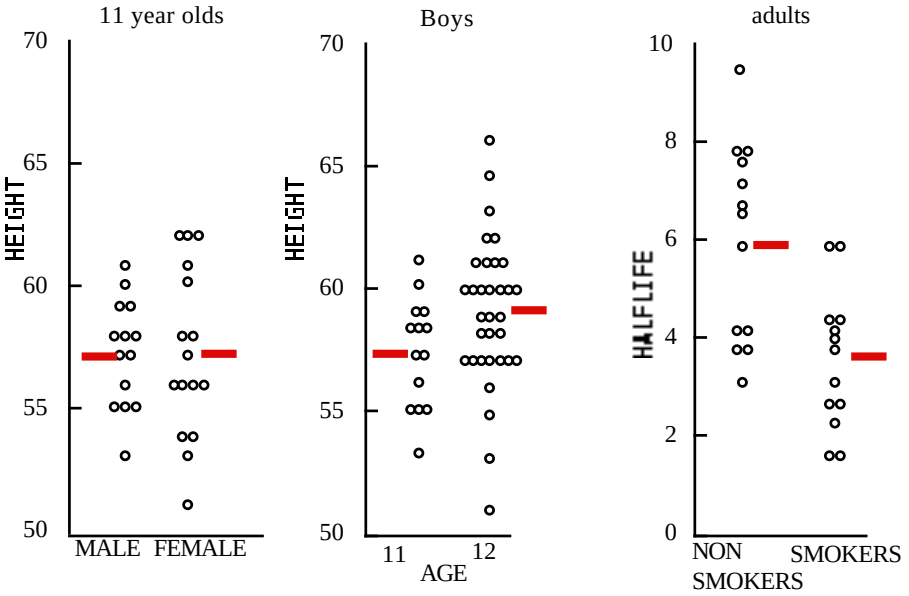
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEPT	1	57.21	0.59	96.2	0.0001

Note the name `SAS` gives to the square root of the average of the squared residuals: Root Mean Square Error, shortened to `ROOT MSE` i.e., average squared deviation = $64.357/13 = 4.95$; $4.95^{1/2} = 2.22$.

Here they are less confusing than SPSS and SYSTAT (to be fair, SEE is used a lot in measurement and psychophysics for variation of measurements on *individuals* [i.e., no $n^{1/2}$ involved], rather than of *statistics*)

6.2 (via Regression) Inference regarding a difference, $\mu_2 - \mu_1$, of 2 means



n	14	16		14	33		13	13	
			$\bar{y} - \bar{y}$			$\bar{y} - \bar{y}$			$\bar{y} - \bar{y}$
$\bar{y}$	57.21	57.25	**0.04**	57.21	59.00	**1.79**	5.95	3.53	**-2.42**
s	2.22	3.41		2.22	3.05		1.99	1.43	

Var*	t	df	Prob	t	df	Prob	t	df	Prob
S	0.03	26.0	0.973	2.24	33.4	0.032	-3.56	21.8	0.002
P	0.03	28	0.974	1.97	45	0.055	-3.56	24	0.002

Var: S = Separate variances t-test; P = Pooled variances t-test

For later: male vs female heights: $s^2_{pooled} = \frac{13 \times 2.22^2 + 15 \times 3.41^2}{13+15} = 8.5 = 2.92^2$.

Statistical Model for difference in mean height of males and females
(see M & M p 663)

Males: $y \sim \mu_{MALE} + \epsilon$ Females: $y \sim \mu_{FEMALE} + \epsilon$

$$\epsilon \sim N(0, \sigma)$$

All: $y \sim \mu_{MALE} + (\mu_{FEMALE} - \mu_{MALE}) \times I_{FEMALE} + \epsilon$

I_{FEMALE} = "Indicator" of Female: so, 0 if Male; 1 if Female.

Or, in more conventional Greek letters... i.e. $\beta_1 = \mu_F - \mu_M$

$$y = \beta_0 + \beta_1 \times I_{FEMALE} + \epsilon.$$

Fitting (i.e. calculating the parameter estimates of) the model

By calculator: $\hat{\beta}_0 = b_1 =$ "slope" $= \sum(x - \bar{x})(y - \bar{y}) / \sum(x - \bar{x})^2$;

$\hat{\beta}_0 = b_0 =$ "intercept" $= \bar{y} - b_1 \times \bar{x}$;

$\hat{\sigma^2} =$ "MSE" $=$ mean[residual^2] $= \sum(y - \hat{y})^2/(n-2)$.

By software: in R: `summary( lm(height  i.female) )`

```
Residuals:
Min 1Q Median 3Q Max
-6.2500 -1.9732 -0.2143 1.7857 4.7500

Coefficients:
 ......... Estimate Std.Error t-value ..Pr(>|t|)
(Intercept) .  57.21 ...  0.78 .  73.22 .  <2e-16
i.female .....  0.04 ...  1.07 ..  0.03 ..  0.974

Residual standard error:  2.924* on 28 df
Mult.  R-Sq:  3.979e-05, Adjusted R-sq:  -0.03567
F-statistic:  0.0011 on 1 & 28 df, p-value:  0.97
```

"Translation"

$\hat{\beta}_0 =$ estimate of $\mu_{MALE} = 0.04$
$\hat{\beta}_1 =$ estimate of $\mu_{FEMALE} - \mu_{MALE} = 57.21$
$\hat{\sigma} = 2.92$.

*Residuals are calculated by squaring the deviation of each y from the estimated (fitted) mean for persons with the same "x" value – in this case those of the same sex – summing them to give 239.357, and dividing this sum by 28 to get 8.5485.

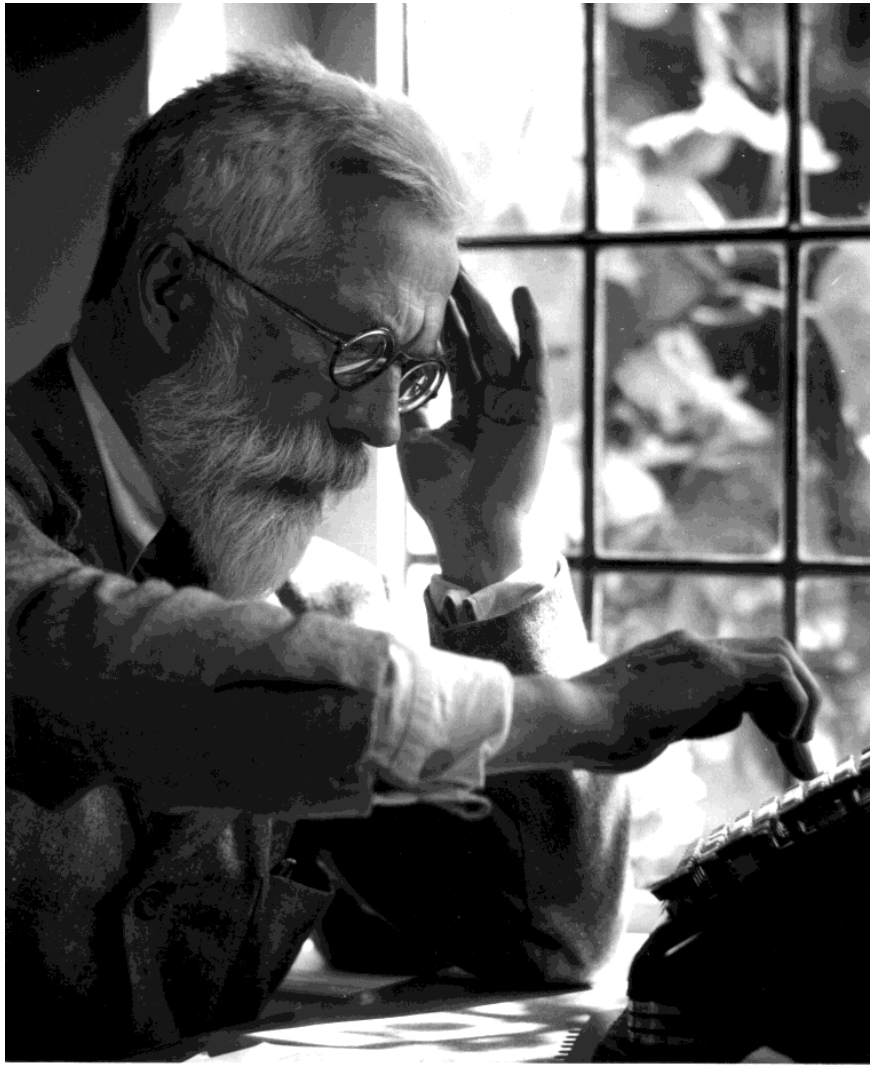
This is the same procedure used to calculate a pooled variance for a 2-sample** t-test! **So the regression model 'recovers' the original means and pooled variance!

Regression approach also reproduces another familiar quantity:

$$\boxed{\text{SE}[\bar{y}_{FEMALE} - \bar{y}_{MALE}] = \text{SE}[\hat{\beta}_{\texttt{i.female}}]}$$

$$\begin{aligned}
\text{SE}[\bar{y}_F - \bar{y}_M] &= (s^2_{pooled})^{1/2} \times (1/n_f + 1/n_m)^{1/2} \\
&= 2.924 \times (1/n_f + 1/n_m)^{1/2} \\
&= 1.07.
\end{aligned}$$

$$\begin{aligned}
\text{SE}[\hat{\beta}_{\texttt{i.f}}] &= \{\text{MSE}/\textstyle\sum(x - \bar{x})^2\}^{1/2} \\
&= \{\text{MSE}/\textstyle\sum(\texttt{i.female} - \overline{\texttt{i.female}})^2\}^{1/2} \\
&= \{\text{MSE}/[(n_f + n_m) \times (n_f/(n_f + n_m)) \times (n_m/(n_f + n_m))]\}^{1/2} \\
&= \{\text{MSE}/[(n_f \times n_f)/(n_f + n_m)]\}^{1/2} \\
&= \{\text{MSE} \times (1/n_f + 1/n_f)\}^{1/2} \\
&= \text{RMSE} \times (1/n_f + 1/n_f)^{1/2} \\
&= 2.924 \times (1/n_f + 1/n_m)^{1/2} \\
&= 1.07.
\end{aligned}$$



FISHER, Sir Ronald Aylmer 1890-1962

Photograph (supplied by Fisher Memorial Committee) by Antony Barrington-Brown, as reproduced as frontispiece of R A Fisher, Collected Papers, Vol.5, Adelaide: Department of Genetics of the University of Adelaide; and also as frontispiece of J F Box, R.A. Fisher: The Life of a Scientist, New York: Wiley 1978.

0 (possible questions for) ASSIGNMENT on mean differences and differences in means

0.1 Questions re van Belle et al.

1. 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 $\mu_2 - \mu_1$.
2. 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)
3. 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]

0.2 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[10]) 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

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

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 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).

0.3 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.

0.4 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.

1. 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."

0.5 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.

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

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

1. 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 ?

2. 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?

0.7 Paracetamol and Fever

1. 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)

2. "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])

3. "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?

4. "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"

5. "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.

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

7. How was the CI calculated?

8. 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]?

9. "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?

10. "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?

0.8 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" [11] [*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.

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.

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

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**
≥ 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 $15 \times 500 = 7500$ children in 15 hospitals and polyclinics randomly assigned to receive the experimental intervention with those in $15 \times 500 = 7500$ children in the 15 assigned to receive the control intervention. Calculate the detectable difference for the IQ and wheezing variables.[12]

0.9 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.[13]

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<.001), 24% more prenatal supplements (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]

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.[14]

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

[12] cf formula in the notes; because of slightly smaller numbers used in this exercise than in grant application, your detectable differences will be slightly different.
[13] Pandey et al. JAMA. 2007;298(16):1867-1875
[14] Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. Int J Epidemiol. 1999; 28(2):319-326.

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.