1 Components of Variance

Researchers are often interested in de-composing observable variation into two or more components or sources. Examples include ...

- quantifying , in **genetic or family studies**, how much of the variation in a quantitative trait (e.g. height, blood pressure, cholesterol) is true between-family variation, how much is true 'between-individual-within-same-family' variation, and how much is real within-individual variation or measurement error. We will examine three such studies.

  In the first (see the 1990 Time magazine story[1]

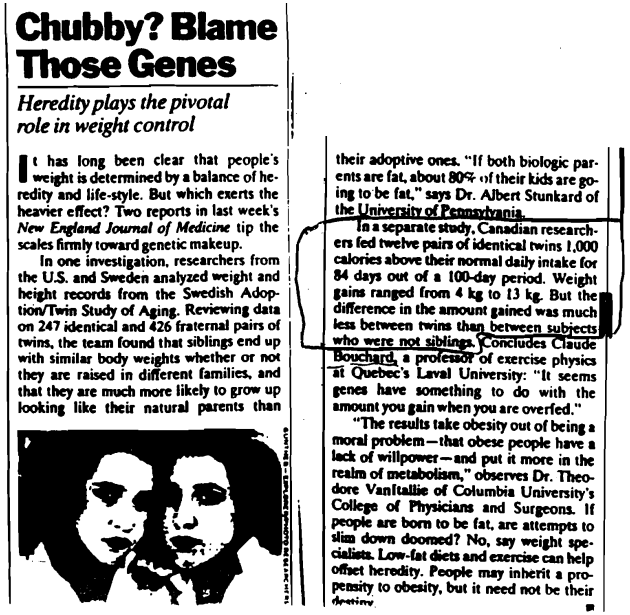
  > Canadian researchers [using U. Laval students as subjects – being in the study was the student's summer job] fed twelve pairs of identical twins 1,000 calories above their normal daily intake for 84 days out of a 100- day period. Weight gains ranged from 4 kg to 13 kg (9 lbs. to 29 lbs.). But the difference in the amount gained was much less between twins than between subjects who were not siblings. Concludes Claude Bouchard, a professor of exercise physics at Quebec's Laval University: "It seems genes have something to do with the amount you gain when you are overfed." Some sets of twins transformed the extra calories into mostly fat, while others converted them into lean muscle.

  The second was motivated by the observation "anatomical, physiological, and epidemiological data indicate that there may be a significant genetic component to prolonged time with and recurrent episodes of otitis media in children". As its objective, it sought

  > to determine the genetic component of time with and episodes of middle ear effusion and acute otitis media (AOM) during the first 2 years of life'.

  The third uses Galton's family stature (height) data to examine between- and within-family differences in adult heights.

[1] "Chubby? Blame those genes: Heredity plays the pivotal role in weight control" `http://www.time.com/time/magazine/article/0,9171,970266,00.html`



Chubby? Blame Those Genes

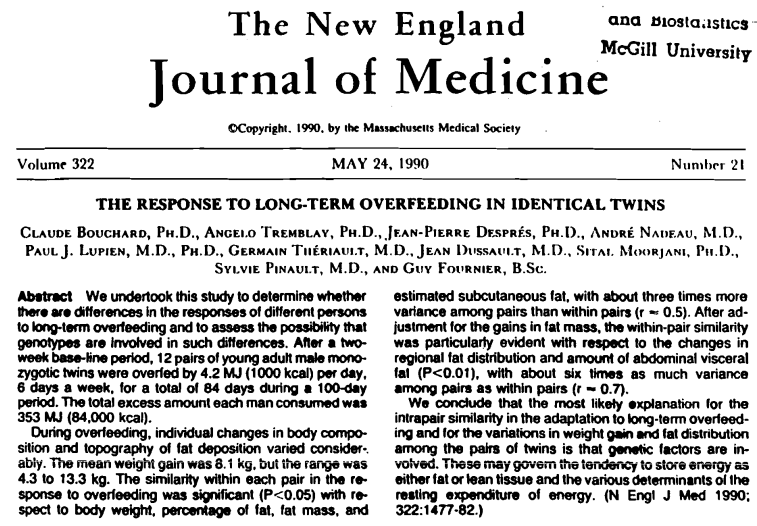
Heredity plays the pivotal role in weight control

It has long been clear that people's weight is determined by a balance of heredity and life-style. But which exerts the heavier effect? Two reports in last week's *New England Journal of Medicine* tip the scales firmly toward genetic makeup.

In one investigation, researchers from the U.S. and Sweden analyzed weight and height records from the Swedish Adoption/Twin Study of Aging. Reviewing data on 247 identical and 426 fraternal pairs of twins, the team found that siblings end up with similar body weights whether or not they are raised in different families, and that they are much more likely to grow up looking like their natural parents than their adoptive ones. "If both biologic parents are fat, about 80% of their kids are going to be fat," says Dr. Albert Stunkard of the University of Pennsylvania.

In a separate study, Canadian researchers fed twelve pairs of identical twins 1,000 calories above their normal daily intake for 84 days out of a 100-day period. Weight gains ranged from 4 kg to 13 kg. But the difference in the amount gained was much less between twins than between subjects who were not siblings. Concludes Claude Bouchard, a professor of exercise physics at Quebec's Laval University: "It seems genes have something to do with the amount you gain when you are overfed."

"The results take obesity out of being a moral problem—that obese people have a lack of willpower—and put it more in the realm of metabolism," observes Dr. Theodore VanItallie of Columbia University's College of Physicians and Surgeons. If people are born to be fat, are attempts to slim down doomed? No, say weight specialists. Low-fat diets and exercise can help offset heredity. People may inherit a propensity to obesity, but it need not be their



The New England Journal of Medicine

Biostatistics McGill University

©Copyright, 1990, by the Massachusetts Medical Society

Volume 322 MAY 24, 1990 Number 21

THE RESPONSE TO LONG-TERM OVERFEEDING IN IDENTICAL TWINS

CLAUDE BOUCHARD, PH.D., ANGELO TREMBLAY, PH.D., JEAN-PIERRE DESPRÉS, PH.D., ANDRÉ NADEAU, M.D., PAUL J. LUPIEN, M.D., PH.D., GERMAIN THÉRIAULT, M.D., JEAN DUSSAULT, M.D., SITAL MOORJANI, PH.D., SYLVIE PINAULT, M.D., AND GUY FOURNIER, B.SC.

Abstract We undertook this study to determine whether there are differences in the responses of different persons to long-term overfeeding and to assess the possibility that genotypes are involved in such differences. After a two-week base-line period, 12 pairs of young adult male monozygotic twins were overfed by 4.2 MJ (1000 kcal) per day, 6 days a week, for a total of 84 days during a 100-day period. The total excess amount each man consumed was 353 MJ (84,000 kcal).

During overfeeding, individual changes in body composition and topography of fat deposition varied considerably. The mean weight gain was 8.1 kg, but the range was 4.3 to 13.3 kg. The similarity within each pair in the response to overfeeding was significant ($P<0.05$) with respect to body weight, percentage of fat, fat mass, and estimated subcutaneous fat, with about three times more variance among pairs than within pairs ($r \approx 0.5$). After adjustment for the gains in fat mass, the within-pair similarity was particularly evident with respect to the changes in regional fat distribution and amount of abdominal visceral fat ($P<0.01$), with about six times as much variance among pairs as within pairs ($r = 0.7$).

We conclude that the most likely explanation for the intrapair similarity in the adaptation to long-term overfeeding and for the variations in weight gain and fat distribution among the pairs of twins is that genetic factors are involved. These may govern the tendency to store energy as either fat or lean tissue and the various determinants of the resting expenditure of energy. (N Engl J Med 1990; 322:1477-82.)

- quantifying, in '**measurement studies**', the amount of measurement error, and expressing it as a coefficient of variation (CV) or reliability coefficient or Intra Class Correlation Coefficient (ICC).

Data analysis

Traditionally, the variance components have been estimated using 'methods of moments' estimators applied to the mean squares calculated in classical ANOVA tables based on a 1-way (or several-way) mixed or random effects model. As statistical computing had become easier, we can now more easily and more flexibly estimate these parameters using a number of approaches and software packages.

But it is best to begin with the classical way. So, following this page, JH has pasted in here 5 pages (numbered 2-6) of orientational material on measurement statistics from a measurement course for physical and occupational therapy students. These students had had limited exposure to statistical concepts in general, and to 'ANOVA' in particular; this lack of familiarity with 'classical' ANOVA[2] does not seem to be limited to such students: many modern 'regression and anova' courses skip the 'anova' altogether, since many of the statistical tests (and anova *tests* were traditionally the focus) can be carried out within a more general regression framework. But in our focus on estimation, and in particular on *variance-estimation*, we have something to learn from the classical anova tables and calculations, and particularly from a concept that is seldom taught within a regression-only course, namely the *Expected Mean Square* or *EMS*. It was mainly used in classical anova to illustrate which Mean Squares should be used in an F test to test which null hypotheses.

Opposite is an excerpt from an older text, showing the EMS for the two simplest '1-way anova' models. We will be more interested in the version where the α's are *random* rather than fixed, but to make it easier, the orientational material starts with the fixed effects model. The anova *calculations* are the same in both the fixed and random-effects models: it is the *use* of the Means-squares that differs in the two models.

[2]By 'classical' I mean the calculations could be easily done by a hand calculator; the data structure was nicely balanced and the data could be laid out in rows and columns, or in a higher-dimensional array, with no missing values, no other explanatory , etc.

ANALYSIS OF VARIANCE AND **EXPECTED MEAN SQUARES** FOR THE ONE-WAY CLASSIFICATION

Model: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}; \quad (i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i; \quad n_\cdot = \sum_i n_i)$

Fixed Effects: $\alpha_1, \ldots \alpha_k$ **fixed** & unknown; $1, 2, \ldots, k$ **exhaustive**; $\sum_i \alpha_i = 0$.

Random Effects: $\alpha_1, \ldots \alpha_k$: **sample** from larger no. of α's, with $\alpha \sim N(0, \sigma_B^2)$

$\epsilon_{ij} \sim N(0, \sigma_W^2)$, *i.i.d.* B: Between; W: Within.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Test Statistic
Between groups	$k-1$	$S_1 = \sum_i \sum_j (\bar{y}_i - \bar{y}_\cdot)^2$	$s_1^2 = \frac{S_1}{k-1}$	$F = \frac{s_1^2}{s_0^2}$
Within groups	$n_\cdot - k$	$S_0 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$	$s_0^2 = \frac{S_0}{n_\cdot - k}$	
Total	$n_\cdot - 1$	$S = \sum_i \sum_j (y_{ij} - \bar{y}_\cdot)^2$		

Source of Variation	Degrees of Freedom	Mean Square	Expected Mean Square (EMS) for model with...	
			Fixed Effects	**Random** Effects
Between groups	$k-1$	s_1^2	$\sigma_W^2 + \frac{\sum_i n_i \alpha_i^2}{k-1}$	$\sigma_W^2 + \frac{1}{k-1}(n_\cdot - \frac{\sum n_i^2}{n_\cdot})\sigma_B^2$ *
Within groups	$n_\cdot - k$	s_0^2	σ_W^2	σ_W^2
Total	$n_\cdot - 1$			

* With equal n's, EMS $= \sigma_W^2 + n\sigma_B^2$; with unequal n's, EMS $> \sigma_W^2 + \bar{n}\sigma_B^2$.

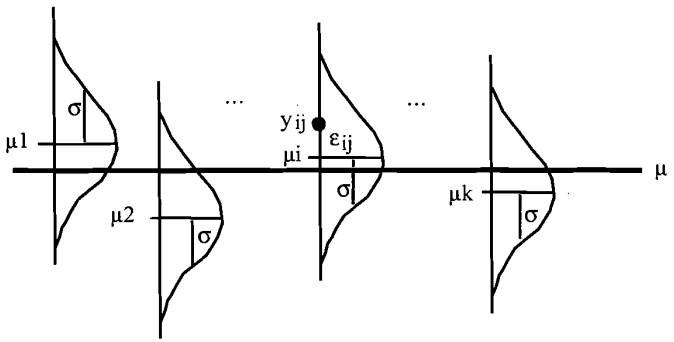
First, a General Orientation to ANOVA and its primary use, namely testing differences between μ's of k (≥2) different groups.

E.g. 1-way ANOVA:

DATA:

	Group					
	1	2	.	i	.	k
Subject						
1	y_{11}	.	.	.	.	.
2						
.	.	.	.	.	.	.
j				y_{ij}		
.	.	.	.	.	.	.
n						y_{kn}
Mean	$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_k$
Variance	s^2_1	s^2_2				s^2_k

MODEL



σ refers to the variation (SD) of all possible individuals in a group;
It is an (unknowable) parameter; it can only be ESTIMATED.

Or, in symbols...

$$y_{ij} = \mu_i + e_{ij} = \mu + (\mu_i - \mu) + e_{ij}$$

DE-COMPOSITION OF OBSERVED (EMPIRICAL) VARIATION

$$\Sigma\Sigma(\bar{y}_{ij} - \bar{y})^2 = \Sigma\Sigma(\bar{y}_i - \bar{y})^2 + \Sigma\Sigma(\bar{y}_{ij} - \bar{y}_i)^2$$

TOTAL Sum of Squares = BETWEEN Groups Sum of Squares + WITHIN Group Sum of Squares

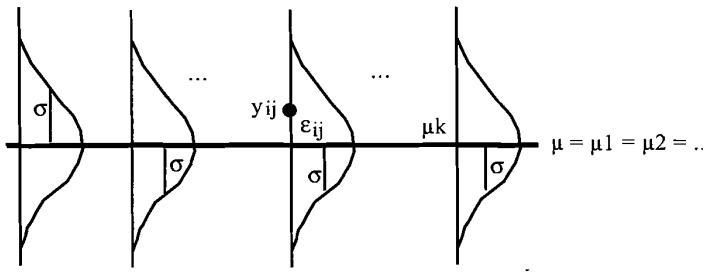
ANOVA TABLE

	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	P-Value
SOURCE	SS	df	MS (= SS /df)	$\frac{MS_{BETWEEN}}{MS_{WITHIN}}$	Prob(>F)
BETWEEN	xx.x	k−1	xx.x	x.xx	0.xx
WITHIN	xx.x	k(n−1)	xx.x		

LOGIC FOR F-TEST (Ratio of variances) as a test of

H_0: $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_k$

UNDER H0



Means, based on samples of n, should vary around μ with a variance of $\frac{\sigma^2}{n}$

Thus, if H_0 is true, and we calculate the empirical variance of the k different $\bar{y}_i$'s, it should give us an unbiased estimate of $\frac{\sigma^2}{n}$

i.e. $\frac{\Sigma[\bar{y}_i - \bar{y}]^2}{k-1}$ is an unbiased estimate of $\frac{\sigma^2}{n}$

i.e. $\frac{n\,\Sigma[\bar{y}_i - \bar{y}]^2}{k-1}$ is an unbiased estimate of σ^2

i.e. $\frac{\Sigma\Sigma[\bar{y}_i - \bar{y}]^2}{k-1} = MS_{BETWEEN}$ is an unbiased estimate of σ^2

Whether or not H_0 is true, the empirical variance of the n (within-group) values y_{i1} to y_{in} i.e. $\frac{\Sigma[\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$ should give us an unbiased estimate of σ^2

i.e. $s^2_i = \frac{\Sigma[\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$ is an unbiased estimate of σ^2

so the average of the k diferent estimates,

$$\frac{1}{k}\sum s^2_i = \frac{1}{k}\sum \frac{\Sigma[\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$$

is also an unbiased estimate of σ^2

i.e. $\frac{\Sigma\Sigma[\bar{y}_{ij} - \bar{y}_i]^2}{k[n-1]} = MS_{WITHIN}$ is an unbiased estimate of σ^2

THUS, under H_0, both $MS_{BETWEEN}$ and MS_{WITHIN} are unbiased estimates of estimates of σ^2 and so their ratio should, apart from sampling variability, be 1. IF however, H_0 is not true, $MS_{BETWEEN}$ will tend to be larger than MS_{WITHIN}, since it contains an extra contribution that is proportional to how far the μ's are from each other.

In this "non-null" case, the $MS_{BETWEEN}$ is an unbiased estimate of

$$\sigma^2 + \frac{\Sigma n[\mu_i - \bar{\mu}]^2}{k-1}$$

and so we expect that, apart from sampling variability, the ratio $\frac{MS_{BETWEEN}}{MS_{WITHIN}}$ should be greater than 1. The tabulated values of the F distribution (tabulated under the assumption that the numerator and denominator of the ratio are both estimaes of the same quantity) can thus be used to assess how extreme the observed F ratio is and to assess the evidence against the H_0 that the μ's are equal.

How ANOVA can be used to estimate Components of Variance used in quantifying Reliability.

The basic ANOVA calculations are the same, but the MODEL underlying them is different. First, in the more common use of ANOVA just described, the groups can be though of as all the levels of the factor of interest. The number of levels is necessarily finite. The groups might be the two genders, all of the age groups, the 4 blood groups, etc. Moreover, when you publish the results, you explicitly identify the groups.

When we come to study subjects, and ask "How big is the intra-subject variation compared with the inter-subject varaition, we will for budget reasons only study a sample of all the possible subjects of interest. We can still number them 1 to k, and we can make n measurements on each subject, so the basic layout of the data doesn'y change. All we do is replace the word 'Group' by 'Subject' and speak of BETWEEN-**SUBJECT** and WITHIN-**SUBJECT** variation. So the data layout is...

DATA:

	Subject					
	1	2	.	i	.	k
Measurement						
1	y_{11}	.	.	.	.	.
2						
.	.	.	.	.	.	.
j				y_{ij}		
.	.	.	.	.	.	.
n						y_{kn}
Mean	$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_k$
Variance	s^2_1	s^2_2				s^2_k

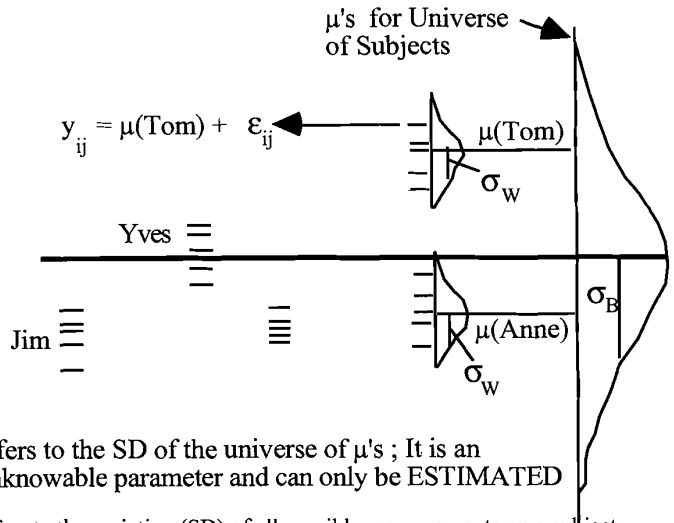
MODEL

The model is different. There is no interest in the specific subjects. Unlike the critical labels "male" anf "female", or "smokers", "nonsmokers" and "exsmokers" to identify groups of interest, we certainly are not going to identify subjects as Yves, Claire, Jean, Anne, Tom, Jim, and Harry in the publication, and nobody would be fussed if in the dataset we used arbitrary subject identifiers to keep track of which measurements were made on whom. we wouldn't even care if the research assistant lost the identities of the subjects -- as long as we know that the correct measurents go with the correct subject!

The "Random Effects" Model uses 2 stages:

(1) random sample of subjects, each with his/her own μ
(2) For each subject, series of random variations around his/her μ

Notice the diagram has considerable 'segregation' of the measurements on different individuals. There is no point in TESTING for (inter-subject) differences in the μ's. The task is rather to estimate the relative magnitudes of the two variance components σ^2_B and σ^2_W.



σ_B refers to the SD of the universe of μ's ; It is an unknowable parameter and can only be ESTIMATED

σ_W refers to the variation (SD) of all possible measurements on a subject It is an (unknowable) parameter; it can only be ESTIMATED.

Or, in symbols...

$$y_{ij} = \mu_i + e_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij}$$
$$= \mu + \alpha_i + \varepsilon_{ij}$$

$\alpha_i \sim N(0, \sigma^2_B)$

$\varepsilon_i \sim N(0, \sigma^2_W)$

DE-COMPOSITION OF OBSERVED (EMPIRICAL) VARIATION

$$\Sigma\Sigma(\bar{y}_{ij} - \bar{y})^2 = \Sigma\Sigma(\bar{y}_i - \bar{y})^2 + \Sigma\Sigma(\bar{y}_{ij} - \bar{y}_i)^2$$

TOTAL Sum of Squares = BETWEEN Subjects Sum of Squares + WITHIN Subjects Sum of Squares

ANOVA TABLE *(Note absence of F and P-value Columns)*

	Sum of Squares	Degrees of Freedom	**Mean Square**	**What the Mean Square is an estimate of***
SOURCE	SS	df	MS (= SS /df)	
BETWEEN Subjects	xx.x	k–1	xx.x	$\sigma^2_W + n\,\sigma^2_B$
WITHIN Subjects	xx.x	k(n–1)	xx.x	σ^2_W

ACTUAL ESTIMATION OF 2 Variance Components

$\mathbf{MS_{BETWEEN}}$ is an unbiased estimate of $\sigma^2_W + n\,\sigma^2_B$

$\mathbf{MS_{WITHIN}}$ is an unbiased estimate of σ^2_W

By subtraction...

$\mathbf{MS_{BETWEEN} - MS_{WITHIN}}$ is an unbiased estimate of $n\,\sigma^2_B$

$\frac{\mathbf{MS_{BETWEEN} - MS_{WITHIN}}}{n}$ is an unbiased estimate of σ^2_B

This is the **definitional** formula; the **computational** formula may be different.

---

* Pardon my ending with a preposition, but I find it difficult to say otherwise. These parameter combinations are also called the "Expected Mean Squares". They are the long-run expectations of the MS statistics As Winston Churchill would say, "For the sake of clarity, this one time this wording is something up which you would put".

Example....

DATA:

	Subject					
	Tom	Anne	Yves	Jean	Claire	
Measurement						
1	4.8	5.5	5.1	6.4	5.8	4.5
2	4.7	5.2	4.9	6.2	6.3	4.1
3	4.9	5.2	5.3	6.6	5.6	4.0
Mean	4.8	5.3	5.1	6.4	5.9	4.2 Variance = 0.614
Variance	0.01	0.03	0.04	0.04	0.13	0.07

ANOVA TABLE (Check... I did it by hand!)

SOURCE	Sum of Squares SS	Degrees of Freedom df	**Mean Square** MS (= SS /df)	**What the Mean Square is an estimate of...** *
BETWEEN Subjects	9.205	5	1.841	$\sigma^2_W + n\,\sigma^2_B$
WITHIN Subjects	0.640	12	0.053	σ^2_W
TOTAL	9.845	17		

ESTIMATES OF VARIANCE COMPONENTS

MS_{WITHIN} = **0.053** is an unbiased estimate of σ^2_W

$\frac{1.841 - 0.053}{3}$ = **0.596** is an unbiased estimate of σ^2_B

1-Way ANOVA Calculations performed by SAS; Components estimated manually

PROC GLM in SAS ==> estimating components 'by hand'

```
DATA a; INPUT  Subject Value; LINES;
1 4.8
1 4.7
...
6 4.5
proc glm; class subject; model value=subject / ss3;
  random subject ;
```

See worked example using earsize data.
If unequal numbers of measurements per subject, see formula in A&B or Fleiss

Estimating Components of Variance using "Black Box"

```
PROC VARCOMP; class subject ; model Value = Subject ;
```

See worked example following...

2 measurements (in mm) of earsize of 8 subjects by each of 4 observers

subject	1				2				3				4			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	67	65	65	64	74	74	74	72	67	68	66	65	65	65	65	65
2nd	67	66	66	66	74	73	71	73	68	67	68	67	64	65	65	64

subject	5				6				7				8			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	65	62	62	61	59	56	55	53	60	62	60	59	66	65	65	63
2nd	61	62	60	61	57	57	57	53	60	65	60	58	66	65	65	65

INTRA-OBSERVER VARIATION (e.g. observer #1)

e.g. observer #1

PROC GLM in SAS ==> estimating components 'by hand'

```
INPUT subject rater occasion earsize; if observer=1;
  The data set has 16 obsns & 4 variables.

proc glm; class subject; model earsize=subject / ss3;
  random subject ;

 General Linear Models Procedure: Class Level Information

Class   Levels Values
SUBJECT    8   1 2 3 4 5 6 7 8 ; # of obsns. in data set = 16

Dependent Variable: EARSIZE
                              Sum of      Mean
Source                  DF  Squares    Square  F Value  Pr > F
Model                    7   341.00     48.71    35.43   0.0001
Error                    8    11.00      1.38
Corrected Total         15   352.00

R-Square      C.V.    Root MSE       EARSIZE Mean
0.968750      1.80    1.17260            65.0

Source                  DF  Type III SS  Mean Square  F Value  Pr > F
SUBJECT                  7   341.00       48.71        35.43    0.0001

Source       Type III Expected Mean Square
SUBJECT      Var(Error) + 2 Var(SUBJECT)

Var(Error) + 2 Var(SUBJECT) = 48.71
Var(Error)                  =  1.38
            2 Var(SUBJECT) = 47.33
              Var(SUBJECT) = 47.33 / 2 = 23.67
```

Estimating Variance components using PROC VARCOMP in SAS

```
proc varcomp; class subject ; model earsize = subject ;

Variance Components Estimation Procedure: Class Level Information

Class    Levels    Values

SUBJECT    8    1 2 3 4 5 6 7 8 ; # obsns in data set = 16

MIVQUE(0) Variance Component Estimation Procedure

                          Estimate
Variance Component        EARSIZE

Var(SUBJECT)               23.67
Var(Error)                  1.38
```

- **ICC** (Fleiss § 1.3)

$$\underline{ICC} = \frac{Var(SUBJECT)}{Var(SUBJECT) + Var(Error)} = \frac{23.67}{23.67 + 1.38} = 0.94$$

1-sided 95% Confidence Interval (see Fleiss p 12)

df for F in CI: (8-1)= 7 and 8

so from Tables of F distribution with 7 & 8 df, F = 3.5

So lower limit of CI for ICC is

$$= \frac{35.43 - 3.5}{35.43 + (2 - 1)\cdot 3.5} = \underline{0.82}$$

EXERCISE: Carry out the estimation procedure for one of the other 3 observers.

Applying this 1-way model to Bouchard's 'chubby genes' data:

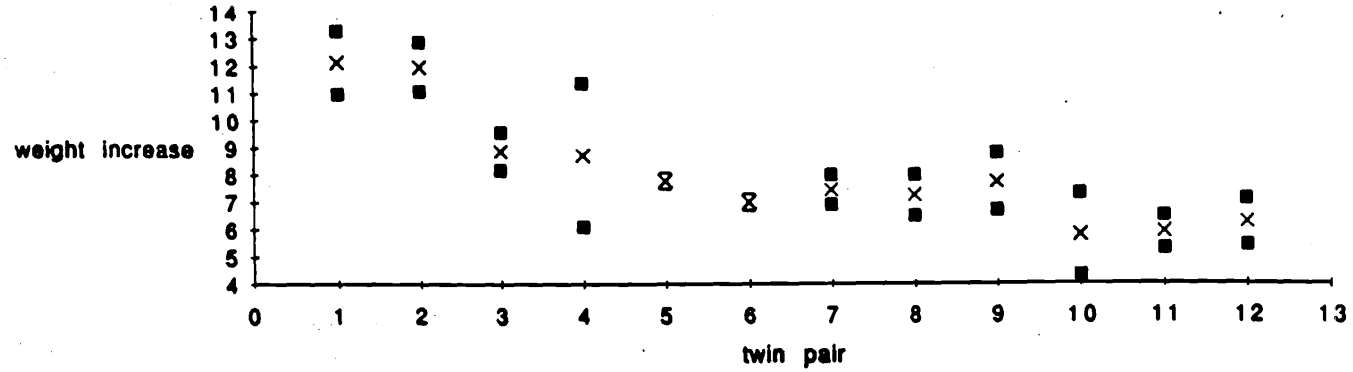
twin pair	1	2	3	4	5	6	7	8	9	10	11	12
μ:	?	?	?	?	?	?	?	?	?	?	?	?
α:	?	?	?	?	?	?	?	?	?	?	?	?
σ:	?	?	?	?	?	?	?	?	?	?	?	?
w. incr. twin A	13.3	11.1	8.2	6.1	7.9	7.1	6.9	6.5	6.7	7.3	6.5	5.4
w. incr. twin B	11	12.9	9.6	11.4	7.7	6.9	8	8	8.8	4.3	5.3	7.1
ybar(i):	12.15	12	8.9	8.75	7.8	7	7.45	7.25	7.75	5.8	5.9	6.25
var(i):	2.645	1.62	0.98	14.05	0.02	0.02	0.605	1.125	2.205	4.5	0.72	1.445
n(i):	2	2	2	2	2	2	2	2	2	2	2	2

ybar: 8.08
var: 5.54
s.d.: 2.35

var(ybars): 4.44
ave(var): 2.49

ANOVA TABLE

Source	Σ Sq	df	Mean Sq	F
B / w Poplns	97.58	11	8.87	3.56
W / n Poplns	29.93	12	2.49	
All	127.51	23	5.54	



From this anova table and the concept of Expected Mean Squares[3], we can, by the method of moments, get estimates of the separate components of that parameter combination.

Quantifying Reliability 3

ICC's (Portnoy and Wilkins)

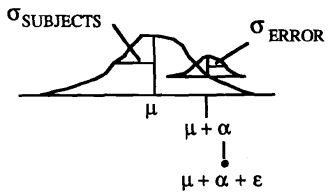
(1) multiple (unlabeled) measurements of each subject

(2) same set of raters measure each subject; raters thought of as a random sample of all possible raters.

(3) as in (2), but these raters studied are the only raters of interest

.....

(1) multiple (unlabeled) measurements of each subject



$$ICC = \frac{\sigma^2_{SUBJECTS}}{\sigma^2_{SUBJECTS} + \sigma^2_{ERROR}}$$

Model for observed data:

$y[\text{subject } i, \text{measurement } j] = \mu + \alpha_i + \varepsilon_{ij}$

EXAMPLE 1

This example is in the spirit of the way the ICC was first used, as a measure of the greater similarity within families than between families: Study by Bouchard (NEJM) on weight gains of 2 members from each of 12 families: It is thought that there will be more variation between members of different families than between members of the same family: family (genes) is though to be a large source of variation; the two twins per family are thought of as 'replicates' from the family and closer to each other (than to others) in their responses. Here the "between" factor is family i.e. families are the subjects and the two twins in the family are just replicates and they don't need to be labeled (if we did label them 1 and 2, the labels would be arbitrary, since the two twins are thought to be 'interchangeable'. (weight gain in Kg over a summer)

model: weight gain for person j in family $i = \mu + \mu + a_i + e_{ij}$

1-way Anova and Expected Mean Square (EMS)

Source	Sum of Sq	d.f	Mean Square	Expected Mean Square
Between (families)	99	11	9.0	$\sigma^2_{\text{"error"}} + k_0 \sigma^2_{\text{between}}$
Error(Within families)	30	12	2.5	$\sigma^2_{\text{"error"}}$
Total	129	23		

In our example, we measure k=2 members from each family, so k_0 is simply 2

[if the k's are unequal, k_0 is somewhat less than the average k... k_0 = average k – (variance of k's) / (n times average k) ...see Fleiss page 10]

Estimation of parameters that go to make up ICC

2.5 is an estimate of $\sigma^2_{\text{"error"}}$

9.0 is an estimate of $\sigma^2_{\text{"error"}} + 2\,\sigma^2_{\text{between}}$

$\therefore$ 6.5 is an estimate of $2\,\sigma^2_{\text{between}}$

$\frac{6.5}{2}$ is an estimate of $\sigma^2_{\text{between}}$

$$\frac{\frac{6.5}{2}}{\frac{6.5}{2} + 2.5} = \frac{3.25}{3.25 + 2.5} = 0.57$$

is an estimate of $ICC = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{error}}$

[3] Think of the EMS for the row in question as the combination of parameters which is (mean-unbiasedly) estimated by the mean square in that row.

You will often find in statistical 'cookbooks' that the ICC, and other concepts – such as those behind the kappa statistic – are defined by the simplest and user-friendliest computational formula its estimator. But biostatisticians should always distinguish between the definition of a parameter and its estimator. Typically the parameter involves Greek letters (some teachers used upper class Roman ones) and the estimator uses data, and the estimate is often denoted by a Greek letter with hat on it, or the lower-case Roman letter equivalent of the upper-case one. In the 'by hand' days, there was the same issue with respect to the definitional formula for a variance or standard deviation versus the user-friendliest computational formula for an estimator of it.

Quantifying Reliability 4

COMPUTATIONAL Formula for "1-way" ICC

$$\frac{\frac{MSbetween - MSwithin}{k_0}}{\frac{MSbetween - MSwithin}{k_0} + MSwithin}$$

$$= \frac{MSbetween - MSwithin}{MSbetween + (k_0 - 1)MSwithin} \quad \text{[shortcut]}$$

is an estimate of the ICC

Notes:
- Streiner and Norman start on page 109 with the 2-way anova for inter-observer variation. There are mistakes in their depiction of the SSerror on p 110 [it should be $(6-6)^2+(4-4)^2+(2-1)^2+...(8-)^2=10$. If one were to do the calculations by hand, one usually calculates the SStotal and then obtains the SSerror by subtraction]
- They then mention the 1-way case, which we have discussed above, as "the observer nested within subject" on page 112
- Fleiss gives methods for calculating CI's for ICC's.

EXAMPLE 2: INTRA-OBSERVER VARIATION FOR 1 OBSERVER

Computations performed on earlier handout...

Var(SUBJECT) = 23.67 Var(ERROR) = 1.38

$\hat{ICC}$ = 23.67 / (23.67 + 1.38) = 0.94

An estimated 94% of observed variation in earsize measurements by this observer is 'real' .. i.e. reflects true between-subject variability.

Note that I say 'an estimated 94% ...". I do this because the 94% is a statistic that is subject to sampling variability (94% is just a point estimate or a 0% Confidence Interval). An interval estimate is given by say a 95% confidence interval for the true ICC (lower bound of a 1-sided CI is 82% ... see previous handout)

Increasing Reliability by averaging several measurements

In 1-way model: $y_{i,j} = \mu + \alpha_i + e_{ij}$

where $var[\alpha_i] = \sigma^2_{\text{between subjects}}$; $var[e_{ij}] = \sigma^2_{\text{"error"}}$

Then if we average k measurements, i.e.,

$$ybar_i = \mu + \alpha_i + ebar_i$$

then

$$\underset{i}{Var}\,[ybar_i] = \sigma^2_{between} + \frac{\sigma^2_{\text{"error"}}}{k}$$

$$\text{So } ICC[k] = \frac{\sigma^2_{between}}{\sigma^2_{between} + \frac{\sigma^2_{\text{"error"}}}{k}}$$

This is called **"Stepped-Up" Reliability**.

≥ 3 components of variance:

when human (or other fallible) raters are involved

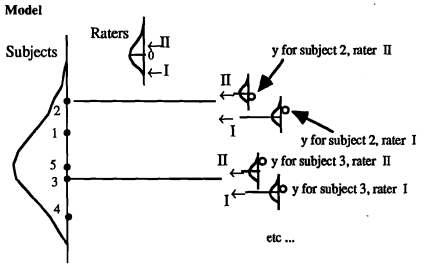
The 64 ear-length measurements below were taken by 4 raters (a subset of the students) who measured 8 subjects (6 other students, as well as the teachers Sharon Wood-Dauphinée and James Hanley) in a (physical and occupational therapy) class on measurement in rehabilitation. The choice of 'objects' measured was prompted by the article Why do old men have big ears? [author James A Heathcote, general practitioner] in the Christmas Edition of the BJM in December 1985, and some follow-up letters the following March – see Resources.

Quantifying Reliability 5

ICC's (Portnoy and Wilkins).

(2) same set of raters measure each subject; raters thought of as a random sample of all possible raters.

- **Model**



$$\mu + \alpha_{[subject]} + \beta_{[rater]} + \varepsilon$$

$$\sigma^2_{subjects} \quad \sigma^2_{raters} \quad \sigma^2_{error}$$

- **From 2- way data layout (subjects x Raters)**

  estimate σ^2"subjects", σ^2"raters" and σ^2"error" by 2-way ANOVA

- **Substitute variance estimates in appropriate ICC form**

e.g. 2 measurements (in mm) of earsize of 8 subjects by each of 4 observers

subject	1				2				3				4			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	67	65	65	64	74	74	74	72	67	68	66	65	65	65	65	65
2nd	67	66	66	66	74	73	71	73	68	67	68	67	64	65	65	64

subject	5				6				7				6			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	65	62	62	61	59	56	55	53	60	62	60	59	66	65	65	63
2nd	61	62	60	61	57	57	57	53	60	65	60	58	66	65	65	65

ESTIMATING INTER-OBSERVER VARIATION from occasion=1;

PROC GLM in SAS ==> estimating components 'by hand'

```
INPUT subject rater occasion earsize; if occasion=1; (32 obsns)

proc glm; class subject rater; model earsize=subject rater / ss3;
  random subject rater;

 General Linear Models Procedure: Class Level Information

Class    Levels    Values
SUBJECT       8    1 2 3 4 5 6 7 8
RATER         4    1 2 3 4 Number of observations in data set = 32

                              Sum of      Mean
Source                  DF  Squares    Square   F Value   Pr > F
Model                   10  764.500    76.45    78.80     0.0001
Error                   21   20.375     0.97
Corrected Total         31  784.875

R-Square       C.V.   Root MSE    EARSIZE Mean
0.974040   1.534577    0.98501        64.1875

Source     DF  Type III SS  Mean Square  F Value  Pr > F
SUBJECT     7   734.875000    104.98      108.20   0.0001
RATER       3    29.625000      9.87       10.18   0.0002

Source       Type III Expected Mean Square
SUBJECT      Var(Error) + 4 Var(SUBJECT)
RATER        Var(Error) + 8 Var(RATER)
```

So... solving 'by hand' for the 3 components...

```
Var(Error) + 4 Var(SUBJECT) = 104.98
Var(Error)                  =   0.97
        ==>  4 Var(SUBJECT) = 104.01
        ==>    Var(SUBJECT) = 104.01 / 4 = 26.00

Var(Error) + 8 Var(RATER)   =   9.87
Var(Error)                  =   0.97
        ==>  8 Var(RATER)   =   8.90
        ==>    Var(RATER)   =   8.90 / 8 = 1.11

               Var(Error)   =   0.97
```

Estimating Variance components using PROC VARCOMP in SAS

```
proc varcomp; class subject rater; model earsize = subject rater;

                          Estimate
Variance Component        EARSIZE
Var(SUBJECT)               26.00
Var(RATER)                  1.11
Var(Error)                  0.97
```

Quantifying Reliability 6

- **ICC: "Raters Random"** (Fleiss § 1.5.2)

$$ICC = \frac{Var(SUBJECT)}{Var(SUBJECT) + Var(RATER) + Var(Error)} = \frac{26.00}{26.00+1.11+0.97} = 0.93$$

1-sided 95% Confidence Interval (see Fleiss p 27)

df for F in CI: (8-1)= 7 and v* , where

$$v^* = \frac{(8-1)(4-1)(4\cdot 0.93\cdot 10.18 + 8[1+(4-1)\cdot 0.93]-4\cdot 0.93)^2}{(8-1)\cdot 4^2\cdot 0.93^2\cdot 10.18^2 + (8[1+(4-1)\cdot 0.93]-4\cdot 0.93)^2} = 8.12$$

so from Tables of F distribution with 7 & 8 df, F = 3.5

So lower limit of CI for ICC is

$$= \frac{8(104.98 - 3.5\cdot 0.97)}{8\cdot 104.98 + 3.5\cdot[4\cdot 9.87 + (8\cdot 4 - 8 - 4)\cdot 0.97]} = 0.78$$

- **ICC: If use one "fixed" observer** (see Fleiss p 23, strategy 3)

$$ICC = \frac{Var(SUBJECT)}{Var(SUBJECT) + Var(Error)} = \frac{26.00}{26.00 + 0.97} = 0.96$$

lower limit of 95% 1-sided CI (eqn 1.49: F = 2.5 ; 7 & 7x3=21 df)

$$ICC = \frac{104.98 - 2.5}{104.98 + (4-1)\cdot 2.5} = 0.91$$

USING ALL THE DATA SIMULTANEOUSLY

(can now estimate subject x Rater interaction .. i.e extent to which raters 'reverse themselves' with different subjects)

Components of variance when use **both** measurements (all 64 obsns)

```
proc varcomp;                          proc varcomp;
 class subject rater;                   class subject rater;
 model earsize = subject rater;         model earsize = subject rater
                                                        subject*rater;
```

Variance Component	Estimate EARSIZE	Variance Component	EARSIZE
Var(SUBJECT)	25.52	Var(SUBJECT)	25.47
Var(RATER)	0.70	Var(RATER)	0.67
Var(Error)	1.37	Var(SUBJECT*RATER)	0.31
		Var(Error)	1.13

m-s exercise 1

i. From first principles, derive the expressions for EMS_B and EMS_W for both the fixed and random effects models in the case of equal n's.

ii. For EMS_B under the random effects model, verify the footnote about the multiplier of σ_B^2 with unequal n's.

Worked examples and use of R/WinBUGS code: cf. Resources

i. Estimation of a log(RateRatio) via (frequentist) Inverse-variance weighting, Likelihood, and Bayesian approaches. Data from article 'Road Trauma in Teenage Male Youth with Childhood Disruptive Behavior Disorders: A Population Based Analysis' by D.A. Redelmeier in PLoS Med 7(11): e1000369. doi:10.1371/journal.pmed.1000369

ii. Estimation of between- and within-family variances (and an icc) from the 'chubby genes' (Bouchard) weight-gain data. See Resources for (a) `R` code to produce the ANOVA table (for method of moments estimation, based on expected mean squares shown in Table on first page of these notes) 'from scratch' [4] i.e., directly from the ANOVA formulae, and to call a 'classical ANOVA' function (b) WinBUGS code for a (Gaussian) random-effects model. `R` code for other (distribution-based) approaches is welcomed.

iii. Estimation of between-subject and between-observer variances (and an icc) using the (64) ear-length measurements collected in the (physical and occupational therapy) class on measurement in rehabilitation.. via the method of moments and via a (Gaussian) random-effects model fitted using WinBUGS. Other approaches are welcomed.

applied exercise 1, option a - otitis media data cf. Resources

i. Derive separate ANOVA tables for the monozygotic and dizygotic twins, and use the method of moments to estimate the components of variance (σ_B^2 and σ_W^2), and the ICC, for each type.

ii. The method of moments approach to variance components estimation does not explicitly use models for the *distribution* of the random effects, or the within-family variations; in addition, the calculation of a confidence intervals for each ICC and the formal statistical comparison of the two ICCs are problematic. Therefore, use an approach[5] that explicitly assumes a Gaussian model for each component of variance, and obtain a point and an interval estimate of (a) each ICC and (b) the ratio of the two ICCs.

[4] GOOGLE origin expression "from scratch"

[5] If using JAGS or WinBUGS with these slightly non-rectangular data, with most families having 2, but some 3, children, you might be able to use an array where one of the dimensions is the maximum of 3, and the 3rd response is set to NA if there are just 2 children. Or you could use the "tall" format, where the data are all in one very long vector, and there is an accompanying vector to say which family it is... the code used in the ear-length data uses this latter (simpler) approach, even though the data in that example has a perfectly 'rectangular' 8 x 4 x 2 array structure.. see the code under Resources.

applied exercise 1, option b - Galton's family data[6] cf. Resources

i. Fit the following *mixed*[7] model for the height of the jth offspring in family i and obtain point and interval estimates for the Between-family variance σ^2_B and the Within-family variance σ^2_W, as well as for $ICC = \sigma^2_B/(\sigma^2_B + \sigma^2_W)$

$$height_{ij} = \mu_{female} + \Delta_{Male} \times I.male_{ij} + b_i + \epsilon_{ij},$$

$$b_i \sim N(0, \sigma_B); \quad \epsilon_{ij} \sim N(0, \sigma_W).$$

You will probably do the fitting via an ML or a Bayesian[8] approach.

Comment on how far you would have been able to get with the method of moments (differences in means squares) approach.

ii. Add what Galton called the 'mid-parent' height (an average of the heights of the 2 parents) as a fixed effect in the above model (technically, σ^2_B and σ^2_W will now have a somewhat different meaning). Interpret the value of the regression[9] coefficient associated with the mid-parent height, and comment on how much (and why) the estimates of σ^2_B and σ^2_W (and the ICC) are affected.

iii. Galton did not use an *additive* model for the male-female height differences; instead he 'transmuted' the female heights by *multiplying* them by a factor, namely the ratio of the mean height of males to that of females:

> "The factor I used was 1.08, which is equivalent to adding a little less than one-twelfth to each female height. It differs slightly from the factors employed by other anthropologists, who, moreover, differ a trifle between themselves; anyhow, it suits my data better than 1.07 or 1.09. I can say confidently that the final result is not of a kind to be sensibly affected by these minute details, because it happened that owing to a mistaken direction, the computer to whom I first entrusted the figures used a somewhat different factor, yet the results came out closely the same."[10]

In a sense, he used a 2-stage estimation process.[11] Suggest how today we might estimate the multiplicative factor from a single-stage regression (*Hint: think of 1.08 as exp[0.077]*).

iv. What model would you suggest to deal with the fact that the SD of height is smaller (by about 8%) for females?

[6] These are taken from the listing found under the Galton tab in JH's website, and thus deliberately omit one family per notebook page. They also include the mis-classification error in his 2004 paper – documented in the notes that accompany that listing. For this exercise, ignore these omissions and the error.

[7] A word about notation: the (2) levels of gender are 'fixed' i.e. they are the only 2 levels possible; their associated regression coefficients (μ_{female} and Δ_{Male}) have meaning and relevance to others and would be identified in any report. The (198) levels of 'family' are 'random' i.e. they are a sample of the effectively infinite number of possible families. Note also the more modern terminology of using the Roman letter b for the random effect and Greek letters αs or βs or Δs for the fixed effects. In the older notation used in ANOVA (cf material at the beginning of this note, and further examples under Resources) it was customary to use Greek letters for both.

[8] If using JAGS or WinBUGS with these non-rectangular data, with different families having different numbers of children, you might be able to use an array where one of the dimensions is the maximum number in any one family, and the height is set to NA if there are fewer than the maximum number of children. Or you could use the "tall" format, where the data are all in one very long vector, and there is an accompanying vector to say which family it is... the code used in the ear-length data uses this latter (simpler) approach, even though the data in that example has a perfectly 'rectangular' 8 x 4 x 2 array structure.. see the code under Resources.

[9] ***You are being part of statistical history here***: When Galton fitted the simple linear regression of offspring height on parental height, he did have a computer (a human one), but he made life easy on himself by using grouped (binned) data and by further reducing the data so he was left with just 9 (x, y) datapoints. He could have applied the Method of Least Squares, developed almost 200 years before, to these 9. But we know that in fact he merely used an "eye" fit, using a "straight edge" to fit his "regression" coefficient of 2/3. This 2/3 became the basis for his description of the phenomenon of "regression to the mean", and the centrepiece of his famous 1886 article "Regression towards mediocrity in hereditary stature". See `http://galton.org/bib/JournalItem.aspx_action=view_id=157` The word "regression" stuck, but our use of it to today has very little to do with its original meaning. In his 3-volume biography of Galton, **Karl Pearson tells us that that 1886 "regression line" was the second such line ever fitted: the first** was the one Galton fitted to the diameters of seeds (sweet peas) in relation to the sizes of their parents, 10 years earlier. Those data, and their analyses, are described in Appendix 1 of his 1886 paper.

[10] See Hanley JA. "Transmuting" Women into Men: Galtons Family Data on Human Stature. The American Statistician, August 2004, Vol. 58, No. 3, page 237. It is available under the r e p r i n t s on JH's website.

[11] He first scaled the heights and then used a simple linear regression on the 'unisex' data. He did not use our type of random effects model. Moreover, when he reduced the unisex data to the 2 way frequency table (1 inch bins for mid-parent height [rows], 1 inch bins for offspring height [columns], with all the offspring in the same mid-parent bin [row] treated as a 'filial array'), he effectively unlinked the offspring from their parents.