# HYPOTHESES TESTED BY F-STATISTICS AVAILABLE IN ANALYSES OF VARIANCE OF UNBALANCED DATA.*

BU-349-M by February, 1971

S. R. Searle
Biometrics Unit, Cornell University, Ithaca, New York
and
Institute of Statistics, Texas A & M University, College Station, Texas

## Abstract

Hypotheses tested by the F-statistics available in familiar analyses of variance of balanced data are well known. However, when data are unbalanced the exact form of comparable hypotheses is not well known. For example, consider the model $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ with $n_{ij}$ observations in the $i^{th}$ row and $j^{th}$ column, there being a rows and b columns in the data. The reduction in sum of squares due to fitting rows after fitting the mean and columns can be symbolized as $R(\alpha|\mu,\beta) = R(\mu,\alpha,\beta) - R(\mu,\beta)$. The F-statistic using this as the numerator sum of squares, with a-1 degrees of freedom, tests the hypothesis

$$H: \left(n_{i\cdot} - \frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i\cdot}}\right)\alpha_i - \sum_{i'\neq i}^{a}\left(\sum_{j=1}^{b}\frac{n_{ij}n_{i'j}}{n_{\cdot j}}\right)\alpha_{i'} + \sum_{j=1}^{b}\left(n_{ij} - \frac{n_{ij}^2}{n_{\cdot j}}\right)\gamma_{ij} - \sum_{i'\neq i}^{a}\sum_{j=1}^{b}\frac{n_{ij}n_{i'j}}{n_{\cdot j}}\gamma_{i'j} = 0$$

for $i = 1, 2, \cdots, a-1$ .

The importance of this and allied results is that in fitting constants to unbalanced data the resulting partitioning of sums of squares (as sometimes summarized in an analysis of variance table) does not provide F-tests of simple, and useful, hypotheses such as H: equality of row effects.

---

* Handout for Symposium on Biomathematics and Computer Science in the Life Sciences, Houston, Texas, March 22-24, 1971.

Paper No. BU-349-M in the Biometrics Unit Mimeograph Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14850.

# HYPOTHESES TESTED BY F-STATISTICS AVAILABLE IN ANALYSES OF VARIANCE OF UNBALANCED DATA.*

BU-349-M by February, 1971

S. R. Searle
Biometrics Unit, Cornell University, Ithaca, New York
and
Institute of Statistics, Texas A & M University, College Station, Texas

Analysis of variance is one of the oldest tools of statistics. Prior to the advent of high-speed computers its use could entail many hours of desk calculator work, but with present-day computing equipment the computational effort can be minimal, and as a result we are seeing more and more calculation of analyses of variance of larger and larger volumes of data.

Analyses of variance can be thought of as coming in one of two forms: of balanced data or of unbalanced data. Balanced data are those in which there is the same number of observations in every subclass of the data. Their usual occurrence is in designed experiments, such as randomized complete blocks, split plots, factorial experiments, and so on. In these situations the corresponding analyses of variance are well known and widely documented, as is also their interpretation.

Unbalanced data are those wherein the different subclasses of the data do not all have the same number of observations but have varying numbers, some of which may be zero---i.e., no observations in some subclasses. This may arise from designed experiments in which, for example, some of the animals died, or some experimental units were lost. More usually, unbalancedness arises with survey

---

* Handout for Symposium on Biomathematics and Computer Science in the Life Sciences, Houston, Texas, March 22-24, 1971.

Paper No. BU-349-M in the Biometrics Unit Mimeograph Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14850.

data, not just data derived from consciously-taken surveys but also data that can be got from currently available experimental units, such as the patients in a hospital or those that come to an out-patient clinic. Illustrative examples are shown in Table 1.

Table 1. Examples of balanced and unbalanced data

Data collected to investigate the effect on basic metabolic rate of 3 different dosages of 4 brands of tranquilizers.

| Brand of Tranquilizer | Numbers of Patients | | | | | |
|---|---|---|---|---|---|---|
| | Balanced Data (designed experiment) | | | Unbalanced Data (bed-patients available in a hospital) | | |
| | Dose Rate | | | Dose Rate | | |
| | 1 | 2 | 3 | 1 | 2 | 3 |
| A | 4 | 4 | 4 | 6 | 0 | 5 |
| B | 4 | 4 | 4 | 0 | 7 | 8 |
| C | 4 | 4 | 4 | 3 | 10 | 4 |
| D | 4 | 4 | 4 | 0 | 5 | 0 |

Analyses of unbalanced data are more difficult to carry out than those of balanced data for three reasons:

(i) Description of the analysis methods for unbalanced data is not as widely available in texts as is that for balanced data---nor are the methods taught so widely.

(ii) The methods for unbalanced data are themselves, when carried out correctly, inherently more difficult than those for balanced data; i.e., the calculations are more complex. For example, in most cases matrix manipulations are required rather than just straightforward summations.

(iii) Even when carried out correctly the results of unbalanced data analyses are more difficult to interpret than are those of balanced data.

The additional difficulty of analyzing unbalanced data over balanced data almost surely means that more errors are made with unbalanced data than with balanced. Errors of calculation may be minimal because a researcher, having obtained correct advice on what to do, will usually go to a computing facility to get his calculations carried out. Assuming that all lines of communication in this procedure operate successfully (ofttimes an unwarranted assumption, unfortunately) the researcher will receive as computer output the correct analysis of his data. Unhappily, interpretation of this output may not always be correct, because interpretation of unbalanced data analyses does not follow "obviously" from that of balanced data. Examples follow.

Table 2. Summary of hypothesis testing in linear models

| | |
|---|---|
| Model: | $\underline{y} = \underline{X}\underline{b} + \underline{e}$, with $\underline{e} \sim N(\underline{0}, \sigma^2\underline{I})$. N observations, with r = rank of $\underline{X}$. |
| Normal equations: | $\underline{X}'\underline{X}\underline{b}^{\circ} = \underline{X}'\underline{y}$ |
| Solution to normal equations: | $\underline{b}^{\circ} = \underline{G}\underline{X}'\underline{y}$, with $\underline{X}'\underline{X}\underline{G}\underline{X}'\underline{X} = \underline{X}'\underline{X}$ |
| Sum of squares due to fitting model: | $R(\underline{b}) = \underline{y}'\underline{X}\underline{G}\underline{X}'\underline{y}$ |
| Residual sum of squares: | $SSE = \underline{y}'\underline{y} - \underline{y}'\underline{X}\underline{G}\underline{X}'\underline{y}$ |
| Estimated variance: | $\hat{\sigma}^2 = SSE/(N - r)$ |
| Hypothesis to be tested: | $H: \underline{K}'\underline{b} = \underline{m}$, with $\underline{K}'$ having full row rank s |
| F-statistic: | $F(H) = (\underline{K}'\underline{b}^{\circ} - \underline{m})'(\underline{K}'\underline{G}\underline{K})^{-1}(\underline{K}'\underline{b}^{\circ} - \underline{m})/s\hat{\sigma}^2$ . |

CASE 1: The simplest model.

Description: The model contains just a mean.

Model: $y_i = \mu + e_i$ (1)

Sum of squares: $R(\mu)$ = sum of squares due to fitting (1)

$= N\bar{y}^2$, where $\bar{y}$ = mean of all y's

Analysis of Variance

| Source | d.f. | Sum of Squares | F-statistic |
|---|---|---|---|
| Mean | 1 | $N\bar{y}^2$ | $F(\mu) = N\bar{y}^2/\hat{\sigma}^2$ |
| Residual | N-1 | $SSE = \underline{y}'\underline{y} - N\bar{y}^2$ | |
| Total | N | $\underline{y}'\underline{y}$ | |

Hypothesis: $F(\mu)$ tests H: $\mu = 0$.

CASE 2: The completely randomized experiment (the 1-way classification).

Description: a classes with $n_i$ observations in $i^{th}$ class

Model: $$y_{ij} = \mu + \alpha_i + e_{ij} \qquad (2)$$

$i = 1, 2, \cdots, a$ and $j = 1, 2, \cdots, n_i$, with $n_. = N$

Sums of squares: $R(\mu)$ = sum of squares due to fitting (1)

$R(\mu,\alpha)$ = sum of squares due to fitting (2)

$$= \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i}$$

Analysis of Variance

| Source | d.f. | Sum of Squares | F-statistic |
|---|---|---|---|
| Mean | 1 | $R(\mu) = N\bar{y}^2$ | $F(\mu) = R(\mu)/\hat{\sigma}^2$ |
| Classes | a-1 | $R(\alpha\|\mu) = R(\mu,\alpha) - R(\mu)$ | $F(\alpha\|\mu) = R(\alpha\|\mu)/(a-1)\hat{\sigma}^2$ |
| Residual | N-a | $SSE = \underline{y}'\underline{y} - R(\mu,\alpha)$ | |
| Total | N | $\underline{y}'\underline{y}$ | |

Hypotheses corresponding to F-statistics

| F-statistic | Unbalanced Data | Balanced Data using $\sum_{i=1}^{a} \alpha_i = 0$ |
|---|---|---|
| $F(\mu)$ | H: $\mu + \Sigma n_i\alpha_i/N = 0$ | H: $\mu = 0$ |
| $F(\alpha\|\mu)$ | H: all $\alpha_i$'s equal | H: all $\alpha_i$'s equal |

CASE 3: Rows and columns (no interaction) -- 2-way classification.

Description: a rows and b columns, 0 or 1 observation in the $i^{th}$ row and $j^{th}$ column.

Model: $$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad (3)$$

$i = 1, 2, \cdots, a, \quad j = 1, 2, \cdots, b, \quad$ and $n_{ij} = 0$ or 1

Sums of squares:

$R(\mu) = N\bar{y}^2$

$R(\mu,\alpha)$ = sum of squares due to fitting (2)

$R(\mu,\beta)$ = sum of squares due to fitting $y_{ij} = \mu + \beta_j + e_{ij}$ analogous to (2)

$R(\mu,\alpha,\beta)$ = sum of squares due to fitting (3)

Analyses of Variance

| Source | d.f. | Sum of Squares | F-statistic |
|---|---|---|---|
| **I: Fitting $\mu$, $\alpha$ after $\mu$, and $\beta$ after $\mu$ and $\alpha$.** | | | |
| Mean, $\mu$ | 1 | $R(\mu) = N\bar{y}^2$ | $F(\mu) = R(\mu)/\hat{\sigma}^2$ |
| $\alpha$ after $\mu$ | a-1 | $R(\alpha\|\mu) = R(\mu,\alpha) - R(\mu)$ | $F(\alpha\|\mu) = R(\alpha\|\mu)/(a-1)\hat{\sigma}^2$ |
| $\beta$ after $\mu$ and $\alpha$ | b-1 | $R(\beta\|\mu,\alpha) = R(\mu,\alpha,\beta) - R(\mu,\alpha)$ | $F(\beta\|\mu,\alpha) = R(\beta\|\mu,\alpha)/(b-1)\hat{\sigma}^2$ |
| Residual | N-a-b+1 | $SSE = \underline{y}'\underline{y} - R(\mu,\alpha,\beta)$ | |
| Total | N | $\underline{y}'\underline{y}$ | |
| **II: Fitting $\mu$, $\beta$ after $\mu$, and $\alpha$ after $\mu$ and $\beta$.** | | | |
| Mean, $\mu$ | 1 | $R(\mu) = N\bar{y}^2$ | $F(\mu) = R(\mu)/\hat{\sigma}^2$ |
| $\beta$ after $\mu$ | b-1 | $R(\beta\|\mu) = R(\mu,\beta) - R(\mu)$ | $F(\beta\|\mu) = R(\beta\|\mu)/(b-1)\hat{\sigma}^2$ |
| $\alpha$ after $\mu$ and $\beta$ | a-1 | $R(\alpha\|\mu,\beta) = R(\mu,\alpha,\beta) - R(\mu,\beta)$ | $F(\alpha\|\mu,\beta) = R(\alpha\|\mu,\beta)/(a-1)\hat{\sigma}^2$ |
| Residual | N-a-b+1 | $SSE = \underline{y}'\underline{y} - R(\mu,\alpha,\beta)$ | |
| Total | N | $\underline{y}'\underline{y}$ | |

Hypotheses corresponding to F-statistics

| F-statistic | Unbalanced Data | Balanced Data using $\Sigma\alpha_i = 0 = \Sigma\beta_j$ |
|---|---|---|
| $F(\mu)$ | H: $\mu + \frac{\Sigma n_{i\cdot}\alpha_i}{N} + \frac{\Sigma n_{\cdot j}\beta_j}{N} = 0$ | H: $\mu = 0$ |
| $F(\alpha\|\mu)$, from I | H: $\alpha_i + \frac{\sum_j n_{ij}\beta_j}{n_{i\cdot}}$ all equal | H: all $\alpha_i$'s equal |
| $F(\alpha\|\mu,\beta)$, from II | H: all $\alpha_i$'s equal | |
| $F(\beta\|\mu)$, from II | H: $\beta_j + \frac{\sum_i n_{ij}\alpha_i}{n_{\cdot j}}$ all equal | H: all $\beta_j$'s equal |
| $F(\beta\|\mu,\alpha)$, from I | H: all $\beta_j$'s equal | |

CASE 4: Rows and columns (with interaction) -- 2-way classification.

Description: a rows, b columns, $n_{ij}$ observations, $n_{ij} \geq 0$.

Model: $$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad (4)$$

$i = 1, 2, \cdots, a;\ j = 1, 2, \cdots, b;\ k = 1, 2, \cdots, n_{ij};$

$n_{ij} \neq 0$ for s subclasses.

Sums of squares:

$R(\mu) = N\bar{y}^2$

$R(\mu,\alpha)$ = sum of squares for $y_{ijk} = \mu + \alpha_i + e_{ijk}$

$R(\mu,\beta)$ = sum of squares for $y_{ijk} = \mu + \beta_j + e_{ijk}$

$R(\mu,\alpha,\beta)$ = sum of squares for $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$

$R(\mu,\alpha,\beta,\gamma)$ = sum of squares due to fitting (4)

Analyses of Variance

| Source | d.f. | Sum of Squares | | F-statistics | |
|---|---|---|---|---|---|
| **I: Fitting $\mu$, $(\alpha\|\mu)$, $(\beta\|\mu,\alpha)$, and $(\gamma\|\mu,\alpha,\beta)$.** | | | | | |
| Mean, $\mu$ | 1 | $R(\mu)$ | $=N\bar{y}^2$ | $F(\mu)$ | $=R(\mu)/\hat{\sigma}^2$ |
| $\alpha$ after $\mu$ | a-1 | $R(\mu\|\alpha)$ | $=R(\mu,\alpha) - R(\mu)$ | $F(\alpha\|\mu)$ | $=R(\alpha\|\mu)/(a-1)\hat{\sigma}^2$ |
| $\beta$ after $\mu$ and $\alpha$ | b-1 | $R(\beta\|\mu,\alpha)$ | $=R(\mu,\alpha,\beta) - R(\mu,\alpha)$ | $F(\beta\|\mu,\alpha)$ | $=R(\beta\|\mu,\alpha)/(b-1)\hat{\sigma}^2$ |
| $\gamma$ after $\mu$, $\alpha$ and $\beta$ | s-a-b+1 | $R(\gamma\|\mu,\alpha,\beta)$ | $=R(\mu,\alpha,\beta,\gamma)-R(\mu,\alpha,\beta)$ | $F(\gamma\|\mu,\alpha,\beta)$ | $=R(\gamma\|\mu,\alpha,\beta)/s'\hat{\sigma}^2$ |
| Residual error | N-s | SSE | $=\underline{y}'\underline{y} - R(\mu,\alpha,\beta,\gamma)$ | (with $s' = s-a-b+1$) | |
| Total | N | | $\underline{y}'\underline{y}$ | | |
| **II. Fitting $\mu$, $(\beta\|\mu)$, $(\alpha\|\mu,\beta)$, and $(\gamma\|\mu,\alpha,\beta)$.** | | | | | |
| Mean, $\mu$ | 1 | $R(\mu)$ | $=N\bar{y}^2$ | $F(\mu)$ as in I. | |
| $\beta$ after $\mu$ | b-1 | $R(\beta\|\mu)$ | $=R(\mu,\beta) - R(\mu)$ | $F(\beta\|\mu)$ | $= R(\beta\|\mu)/(b-1)\hat{\sigma}^2$ |
| $\alpha$ after $\mu$ and $\beta$ | a-1 | $R(\alpha\|\mu,\beta)$ | $=R(\mu,\alpha,\beta) - R(\mu,\beta)$ | $F(\alpha\|\mu,\beta)$ | $= R(\alpha\|\mu,\beta)/(a-1)\hat{\sigma}^2$ |
| $\gamma$ after $\mu$, $\alpha$ and $\beta$ | s-a-b+1 | $R(\gamma\|\mu,\alpha,\beta)$ | as in I. | $F(\gamma\|\mu,\alpha,\beta)$ | as in I. |
| Residual error | N-s | SSE | as in I. | | |
| Total | N | | $\underline{y}'\underline{y}$ | | |

Hypotheses corresponding to F-statistics

| F-statistic | Unbalanced Data | Balanced Data using $\Sigma\alpha_i=0$, $\Sigma\beta_j=0$, $\Sigma\gamma_{ij}=0$ for all i and j; $\Sigma_i\gamma_{ij}=0$ for all j. |
|---|---|---|
| $F(\mu)$ | H: $\mu + \Sigma_i n_{i\cdot}\alpha_i/N + \Sigma n_{\cdot j}\beta_j/N + \Sigma\Sigma n_{ij}\gamma_{ij}/N = 0$ | H: $\mu = 0$ |
| $F(\alpha\|\mu)$, from I | H: $\alpha_i + \frac{\Sigma_j n_{ij}(\beta_j + \gamma_{ij})}{n_{i\cdot}}$ all equal | H: all $\alpha_i$'s equal |
| $F(\alpha\|\mu,\beta)$, from II | H: $(n_{i\cdot} - \Sigma_j \frac{n_{ij}^2}{n_{\cdot j}})\alpha_i - \Sigma_{i'\neq i}(\Sigma_j \frac{n_{ij}n_{i'j}}{n_{\cdot j}})\alpha_{i'} + \Sigma_j(n_{ij} - n_{ij}^2/n_{\cdot j})\gamma_{ij} - \Sigma_{i'\neq i}\Sigma_j(n_{ij}n_{i'j}/n_{\cdot j})\gamma_{i'j} = 0$ for $i = 1, 2, \cdots, a-1$ | |
| $F(\beta\|\mu)$, from II | H: $\beta_j + \frac{\Sigma_i n_{ij}(\alpha_i + \gamma_{ij})}{n_{\cdot j}}$ all equal | H: all $\beta_j$'s equal |
| $F(\beta\|\alpha,\mu)$, from I | H: $(n_{\cdot j} - \Sigma_i \frac{n_{ij}^2}{n_{i\cdot}})\beta_j - \Sigma_{j'\neq j}(\Sigma_i \frac{n_{ij}n_{ij'}}{n_{i\cdot}})\beta_{j'} + \Sigma_i(n_{ij} - n_{ij}^2/n_{i\cdot})\gamma_{ij} - \Sigma_{j'\neq j}\Sigma_i(n_{ij}n_{ij'}/n_{i\cdot})\gamma_{ij} = 0$ for $j = 1, 2, \cdots, b-1$ | |
| $F(\gamma\|\mu,\alpha,\beta)$ | H: Any column vector of s-a-b+1 linearly independent functions of $\theta_{ij,i'j'} = \gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'}$ where the $\gamma$'s in such functions are all from cells that contain observations. $= 0$ | H: all $\gamma_{ij}$ equal |

CASE 5: 3-way classification (no interactions)

Description: a classes, b rows, and c columns.

Model: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}$

6 Partitionings of total sum of squares, $\underline{y}'\underline{y}$

| (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|
| $R(\mu)$ | $R(\mu)$ | $R(\mu)$ | $R(\mu)$ | $R(\mu)$ | $R(\mu)$ |
| $R(\alpha\|\mu)$ | $R(\alpha\|\mu)$ | $R(\beta\|\mu)$ | $R(\beta\|\mu)$ | $R(\gamma\|\mu)$ | $R(\gamma\|\mu)$ |
| $R(\beta\|\mu,\alpha)$ | $R(\gamma\|\mu,\alpha)$ | $R(\alpha\|\mu,\beta)$ | $R(\gamma\|\mu,\beta)$ | $R(\alpha\|\mu,\gamma)$ | $R(\beta\|\mu,\gamma)$ |
| $R(\gamma\|\mu,\alpha,\beta)$ | $R(\beta\|\mu,\alpha,\gamma)$ | $R(\gamma\|\mu,\alpha,\beta)$ | $R(\alpha\|\mu,\beta,\gamma)$ | $R(\beta\|\mu,\alpha,\gamma)$ | $R(\alpha\|\mu,\beta,\gamma)$ |
| SSE | SSE | SSE | SSE | SSE | SSE |
| $\underline{y}'\underline{y}$ | $\underline{y}'\underline{y}$ | $\underline{y}'\underline{y}$ | $\underline{y}'\underline{y}$ | $\underline{y}'\underline{y}$ | $\underline{y}'\underline{y}$ |