University of Washington

STATISTICS

# Elements of Statistical Methods
# The Analysis of Variance (ANOVA) (Ch 12)

Fritz Scholz

Spring Quarter 2010

May 20, 2010

# The Basic k-Sample Problem

Here we generalize the 2-sample problem to the $k$-sample problem.

Because of increased complexity we also make certain simplifying assumptions.

We have $k$ independent random samples of respective sizes $n_1, \ldots, n_k$ from populations with distributions $P_1, \ldots, P_k$.

It is convenient to express this in double index notation

$$\begin{array}{rcl} X_{11}, \ldots, X_{1n_1} & \sim & P_1 \\ X_{21}, \ldots, X_{2n_2} & \sim & P_2 \\ \ldots & \vdots & \ldots \\ X_{k1}, \ldots, X_{kn_k} & \sim & P_k \end{array}$$

Succinctly we express this as $X_{ij} \sim P_i$.

This situation occurs when comparing several different treatments or methods, or when trying to assess whether samples can be pooled or not.

# Basic Assumptions

We assume the following:

1. The $X_{ij} \sim P_i$ are all independent continuous random variables.

2. $P_i$ has location parameter $\theta_i$, e.g., $\theta_i = \mu_i = EX_{ij}$ or $\theta_i = q_2(X_{ij})$.

3. We observe random samples $\vec{x}_i = (x_{i1}, \ldots, x_{in_i})$, $i = 1, \ldots, k$, from which we want to draw inferences about $\theta_1, \ldots, \theta_k$.

4. $P_i = \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, k$

Note the normality and the common variance assumption.

Both can be overcome, with more complications than we wish to face in this course.

# The Fundamental Null Hypothesis

We test the null hypothesis

$H_0 : \mu_1 = \ldots = \mu_k$ against the alternative $H_1$ : not all $\mu_i$ are the same.

Let $$N = \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \sum_{j=1}^{n_i} 1$$ and $$\bar{\mu}. = \sum_{i=1}^{k} \frac{n_i}{N} \mu_i = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mu_i$$

where $\bar{\mu}.$ is called the population grand mean or simply the grand mean.

It is the weighted average (weights $n_i/N$) of the individual means.

$$\sum_{i=1}^{k} \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^{k} n_i = \frac{N}{N} = 1$$

When all the means are the same, say $= \mu$, then

$$\bar{\mu}. = \sum_{i=1}^{k} (n_i/N)\mu = \mu \sum_{i=1}^{k} (n_i/N) = \mu$$

# Discrepancy Between the Means $\mu_1, \ldots, \mu_k$

The discrepancy between the means $\mu_1, \ldots, \mu_k$ is most commonly expressed as

$$\gamma = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mu_i - \bar{\mu}.)^2 = \sum_{i=1}^{k} n_i (\mu_i - \bar{\mu}.)^2 \quad \text{with} \quad \gamma = 0 \iff \mu_1 = \ldots = \mu_k$$

Thus our previous testing problem becomes: $H'_0 : \gamma = 0$ versus $H'_1 : \gamma > 0$

Note that $H_0$ or $H'_0$ mean that all $k$ sampled distribution are the same

normal distribution, since a common $\sigma$ and normality was assumed a priori.

# Using the Plug-In Principle to Estimate $\gamma$

The plug-in principle suggests to estimate the individual $\mu_i$ by

$$\bar{X}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad \text{the } i^{\text{th}} \text{ sample mean}$$

and the grand mean $\bar{\mu}.$ by using the sample grand mean

$$\bar{X}_{..} = \sum_{i=1}^{k} \frac{n_i}{N} \bar{X}_{i.} = \sum_{i=1}^{k} \frac{n_i}{N} \left( \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \right) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

$\bar{X}_{i.}$ and $\bar{X}_{..}$ are unbiased estimators of $\mu_i$ and $\bar{\mu}.$, respectively.

The natural (plug-in) estimator of $\gamma$ is

$$SS_B = \sum_{i=1}^{k} n_i (\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})^2 \quad = \text{Between groups Sum of Squares}$$

On intuitive grounds we should reject $H_0$ when $SS_B$ is sufficiently large.

We will distinguish two cases: $\sigma^2$ known and $\sigma^2$ unknown.

# Testing $H_0$ when $\sigma^2$ is Known

**Theorem:** Under $H_0$ and the assumption of normality and common variance $\sigma^2$ we have

$$\frac{SS_B}{\sigma^2} \sim \chi^2(k-1)$$

This theorem motivates the use of $SS_B$ instead of other discrepancy metrics, such as

$$\sum_{i=1}^{k} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \qquad \text{or} \qquad \max_{i=1,\ldots,k} \{|\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}|\}$$

This distributional result provides a reference or null distribution under $H_0$ against which to compare values of $SS_B$ that are possibly too large.

We can use $\texttt{qchisq}(\ldots,\texttt{k}-1)$ or $1-\texttt{pchisq}(\ldots,\texttt{k}-1)$ to obtain appropriate critical values or significance probabilities.

# Example when $\sigma^2$ is Known

Suppose we have $n_1 = 20, n_2 = 25$, and $n_3 = 30$ observations with respective sample means $\bar{x}_1 = 1.489$, $\bar{x}_2 = 1.712$ and $\bar{x}_3 = 3.082$.

Assume a known variance $\sigma^2 = 9$ when testing $H_0$ against $H_1$.

As sample grand mean we get `(20*1.489+25*1.712+30*3.082)/75=2.200533`

and thus for $ss_B$ we get

```
> 20*(1.489-2.200533)^2+25*(1.712-2.200533)^2+30*(3.082-2.200533)^2
[1] 39.40172
```

The text uses an alternate formula for $ss_B$ which is mathematically equivalent, but can lead to numerical significance loss.

$$ss_B = \sum_{i=1}^{k} n_i \bar{x}_{i\cdot}^2 - \frac{1}{N}\left(\sum_{i=1}^{k} n_i \bar{x}_{i\cdot}\right)^2$$

difference of possibly large squares

# Example (continued)

For significance level $\alpha = 0.05$ we should reject $H_0$ when

$$39.40172/9 = 4.377969 \geq \texttt{qchisq}(0.95, 2) = 5.991465$$

which is not the case, i.e., the result is not significant at level $\alpha = 0.05$.

As significance probability we get

$$\begin{aligned} \mathbf{p}(ss_B/\sigma^2) &= P_{H_0}(SS_B/\sigma^2 \geq ss_B/\sigma^2) \\ &= 1 - \texttt{pchisq}(4.377969, 2) = 0.1120305 \end{aligned}$$

confirming the previous conclusion since $0.1120305 > 0.05$.

# Unknown Population Variance $\sigma^2$

As in the 2-sample case we make use of the following facts:

$$\frac{(n_i-1)S_i^2}{\sigma^2} = \frac{\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2}{\sigma^2} \sim \chi^2(n_i-1) \quad \text{independently for } i=1,\ldots,k$$

From our results concerning the sum of independent $\chi^2$ random variables we get

$$\frac{(n_1-1)S_1^2}{\sigma^2}+\ldots+\frac{(n_k-1)S_k^2}{\sigma^2} = \frac{(n_1-1)S_1^2+\ldots+(n_k-1)S_k^2}{\sigma^2}$$

$$\sim \chi^2((n_1-1)+\ldots+(n_k-1)) = \chi^2(N-k) \quad \text{with expectation } N-k$$

$$S_P^2 = \sigma^2\,\frac{(n_1-1)S_1^2+\ldots+(n_k-1)S_k^2}{\sigma^2(N-k)} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)S_i^2$$

$$= \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2 \quad \text{with expectation } \sigma^2$$

$$\Longrightarrow S_P^2 \quad \text{is an unbiased estimator of } \sigma^2 \quad \text{and} \quad \frac{(N-k)\,S_P^2}{\sigma^2} \sim \chi^2(N-k)$$

# Sum of Squares Decomposition

We call

$$SS_W = (N-k)\, S_P^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2$$

the within group or error sum of squares. We also introduce

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{\cdot\cdot})^2$$

as the total sum of squares.

We have the following sum of squares decomposition

$$SS_T = SS_W + SS_B$$

or

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

which is a form of the Pythagorean Theorem $c^2 = a^2 + b^2$ in a right triangle.

# Orthogonality and Pythagorean Theorem

Think of the

$$X_{ij} - \bar{X}_{\cdot\cdot} = (X_{ij} - \bar{X}_{i\cdot}) + (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) \quad j = 1,\ldots,n_i, i = 1,\ldots,k$$

on the left side of $=$ as one long vector with $N$ components, expressed as the sum of two orthogonal vectors (on the right side) of same length $N$.

Orthogonality of the latter comes from

$$\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})=0 \Longrightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})=\sum_{i=1}^{k}(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})=0$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{\cdot\cdot})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2$$

since $(a+b)^2 = a^2+b^2+2ab$ with $a = X_{ij} - \bar{X}_{i\cdot}$ and $b = \bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}$

$$(X_{ij}-\bar{X}_{\cdot\cdot})^2 = (X_{ij}-\bar{X}_{i\cdot})^2 + (\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2 + 2(X_{ij}-\bar{X}_{i\cdot})(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})$$

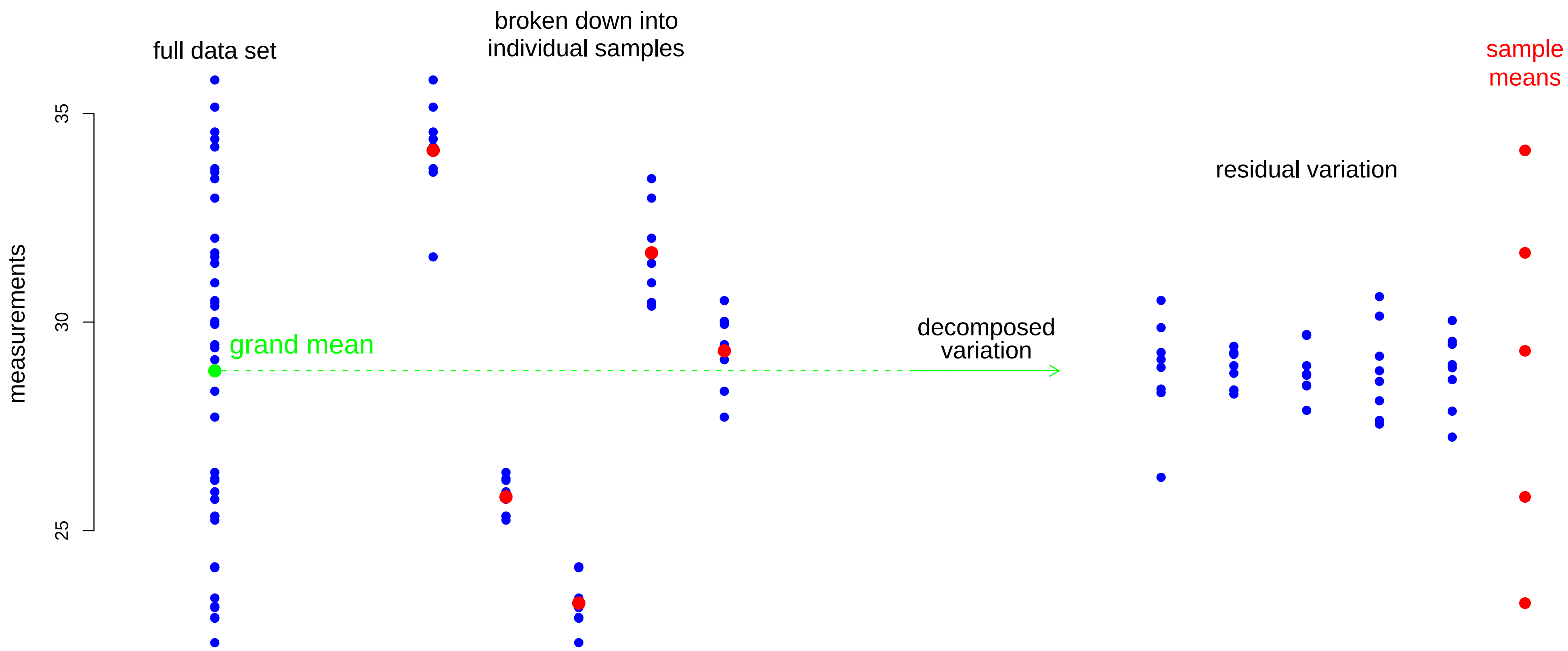
and the double summation of the terms $2(X_{ij}-\bar{X}_{i\cdot})(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})$ is zero.

# Sum of Squares Decomposition and ANOVA

The previous sum of square decomposition breaks down the variability of the data around the sample grand mean into the variability within each group (summed over all $k$ groups) and into variability of the group centers (sample means).

This differentiated view of the variability (expressed via sums of squares) is the essence of the Analysis of Variance (ANOVA).

# Graphical Illustration of ANOVA



# Distributional Facts

**Theorem:** Assuming $k$ independent normal random samples of respective sizes $n_1, \ldots, n_k$, with same variance $\sigma^2$ but possibly different means $\mu_1, \ldots, \mu_k$, we have

$$(N-k)S_P^2/\sigma^2 = SS_W/\sigma^2 \sim \chi^2(N-k) \quad \text{(claimed previously)}$$

and $SS_W$ and $SS_B$ are independent.

If in addition $H_0 : \mu_1 = \ldots = \mu_k$ holds, then

$$SS_T/\sigma^2 \sim \chi^2(N-1) \qquad \text{and} \qquad SS_B/\sigma^2 \sim \chi^2(k-1) \quad \text{(claimed previously)}$$

Comment: The independence of $SS_W$ and $SS_B$ follows from the independence of $\bar{X}_{i\cdot}$ and $S_i^2$ for $i = 1, \ldots, k$ and the independence of the samples.

Note that $SS_W$ is an aggregate of the $S_i^2$ and $SS_B$ is computed solely from the $\bar{X}_{i\cdot}$.

# The $F$-Test

Since our previous test statistic (for known variance $\sigma^2$) was $SS_B/\sigma^2$, it would seem natural to use, in the case of unknown variance, the statistic $SS_B/S_P^2$, i.e., replace the unknown $\sigma^2$ by its unbiased estimator $S_P^2$.

We would reject $H_0$ when $SS_B/S_P^2$ is too large.

However, to link up to a known and standard distribution we use

$$F = \frac{SS_B/(k-1)}{SS_W/(N-k)} = \frac{1}{k-1}\frac{SS_B}{S_P^2} \quad \text{and reject } H_0 \text{ when } F \text{ is too large}$$

**Corollary:** Under the normality assumption with same variance $\sigma^2$ and $H_0 : \mu_1 = \ldots = \mu_k$ we have

$$F = \frac{SS_B/(k-1)}{SS_W/(N-k)} = \frac{(SS_B/\sigma^2)/(k-1)}{(SS_W/\sigma^2)/(N-k)} \sim F(k-1, N-k)$$

which is the $F$-distribution with $k-1$ and $N-k$ degrees of freedom, respectively.

Immediate consequence of the previous theorem and the $F$ distribution definition.

# Rationale for the $F$-Test

When $H_0 : \mu_1 = \ldots = \mu_k$ is not true, it will result in the sample averages $\bar{X}_{1\cdot}, \ldots, \bar{X}_{k\cdot}$ being more dispersed.

Thus $SS_B$ tends to be larger under $H_1$ than under $H_0 : \mu_1 = \ldots = \mu_k$.

The behavior of the denominator is not affected by $H_0$ true or false.
It always is an unbiased estimator of $\sigma^2$.

Thus we will expect to see larger values of $F$ under $H_1$ than under $H_0$.

Unusually large values of $F$ should be compared with the null distribution of $F$.

Do this by using critical values for given $\alpha$ or via significance probabilities.

# An Example

| | $i=1$ | $i=2$ | $i=3$ |
|---|---|---|---|
| $n_i$ | 25 | 20 | 20 |
| $\bar{x}_{i\cdot}$ | 9.783685 | 10.908170 | 15.002820 |
| $s_i^2$ | 29.89214 | 18.75800 | 51.41654 |

ANOVA Table

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares | Test Statistic | Significance Probability |
|---|---|---|---|---|---|
| Between | $SS_B$ | $k-1$ | $MS_B = \frac{SS_B}{k-1}$ | $F = \frac{MS_B}{MS_W}$ | $\mathbf{p}(f)$ |
| Within | $SS_W$ | $N-k$ | $MS_W = \frac{SS_W}{N-k} = S_P^2$ | | |
| Total | $SS_T$ | $N-1$ | | | |

| Source | $SS$ | df | $MS$ | $F$ | $\mathbf{p}$ |
|---|---|---|---|---|---|
| Between | 322.4366 | 2 | 161.21832 | 4.87414117 | 0.01081398 |
| Within | 2050.7276 | 62 | 33.07625 | | |
| Total | 2373.1643 | 64 | | | |

$R^2 = SS_B/SS_T = 0.136$ proportion of $SS_T$ “explained” by the $\bar{x}_{i\cdot}$ variation.

# Code for Previous Example

```
anova12.2 <- function(alpha=.05){
n <- c(25, 20, 20)
xbar <- c(9.783685, 10.908170, 15.002820)
s2 <- c(29.89214, 18.75800,  51.41654)
N <- sum(n); k <- length(n)
ssW <- sum((n-1)*s2); xbar.grand <- sum(n*xbar/N)
ssB <- sum(n*(xbar-xbar.grand)^2)
F.stat <- (ssB/(k-1))/(ssW/(N-k)); F.crit <- qf(1-alpha,k-1,N-k)
pval <- 1-pf(F.stat,k-1,N-k)
ssT <- ssW+ssB; ss <- c(ssB,ssW,ssT); ms <- c(ssB/(k-1),ssW/(N-k))
stats <- c(F.stat,F.crit,pval)
names(ss) <- c("ssB","ssW","ssT")
names(ms) <- c("msB","msW")
names(stats) <- c("F.observed","F.crit","p-value")
list(ss=ss,ms=ms,stats=stats)}
```

# Output from `anova12.2`

```
> anova12.2(alpha=0.05)
$ss
      ssB       ssW       ssT
 322.4366 2050.7276 2373.1643

$ms
      msB       msW
161.21832  33.07625

$stats
F.observed     F.crit    p-value
4.87414117 3.14525838 0.01081398
```

# Surface Insulation Resistance (SIR)

Circuit boards in aircraft experience intermittant failures due to insulation problems caused by residual solder flux. Different fluxes, X,Y,Z, are investigated.

http://en.wikipedia.org/wiki/Flux_(metallurgy)

| SIR | FLUX |
|---|---|
| 9.9 | X |
| 9.6 | X |
| 9.6 | X |
| 9.7 | X |
| 9.5 | X |
| 10.0 | X |
| 10.7 | Y |
| 10.4 | Y |
| 9.5 | Y |
| 9.6 | Y |
| 9.8 | Y |
| 9.9 | Y |
| 10.9 | Z |
| 11.0 | Z |
| 9.5 | Z |
| 10.0 | Z |
| 11.7 | Z |
| 10.2 | Z |

SIRFLUX.csv file available under lectures.

```
> SIRFLUX <- read.csv("SIRFLUX.csv",header=T)
> names(SIRFLUX)
[1] "SIR"  "FLUX"
> anova(lm(SIR~FLUX,data=SIRFLUX))
Analysis of Variance Table

Response: SIR
          Df Sum Sq Mean Sq F value  Pr(>F)
FLUX       2 2.1733 1.08667  3.6452 0.05126 .
Residuals 15 4.4717 0.29811
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

# `anova(lm(SIR~FLUX,data=SIRFLUX))`

`SIRFLUX <- read.csv("SIRFLUX.csv",header=T)` reads the csv file into a data frame called `SIRFLUX`.

`lm(SIR~FLUX,data=SIRFLUX)` does the ANOVA calculations,

differentiating the responses `SIR` by the `FLUX` variable (factor).

It knows the meaning of `SIR` and `FLUX` via the `data=SIRFLUX` specification.

The command `anova(lm(...))` just creates the nicely formatted ANOVA table

output from the analysis performed by `lm(SIR~FLUX,data=SIRFLUX)`