USING STATISTICS IN RESEARCH

David A. Stephens

Department of Mathematics, Imperial College

d.stephens@imperial.ac.uk
stats.ma.ic.ac.uk/~das01/StatsShortCourse/

June 25, 2003

WEEK 4

REPEATED MEASURES

&

MULTIVARIATE RESPONSES

SECTION 7.

REPEATED MEASURES AND MULTIVARIATE RESPONSE

A typical experimental design involves taking repeated measurements of the **same** feature on the **same** experimental unit.

- observations made at time $t = 0, t_1, t_2, ...$
- observations made on treatment 1,2,3,...

The resulting observed values are necessarily **correlated** through "time" in this design because of the "within individual" factors. However, we may still be interested in the impact of different **predictors/covariates/fixed effects.**

7.1 MODELLING ASSUMPTIONS

We want to assess the systematic differences in **mean** response level (between time points, between treatment groups etc.) whilst accounting for the correlation in the observed data. Let y_{ijk} be the

- k^{th} observation for the
- j^{th} experimental unit in the
- i^{th} group

for $i = 1, ..., n$, $j = 1, ..., n_i$ and $k = 1, ..., K_{ij}$. Then

$$E\left[Y_{ijk}\right] = \mu_{ijk}$$

and we attempt to model μ_{ijk}, but $\mathbf{Y}_{ij} = \left(Y_{ij1}, ...Y_{ijK_{ij}}\right)$ have some **multivariate probability distribution** due to the within individual correlation.

A common (and typical) assumption is to use a multivariate Normal distribution,

$$\mathbf{Y}_{ij} \sim N_{K_{ij}} \left(\boldsymbol{\mu}_{ij}, \Sigma_{ij} \right)$$

where

- $\boldsymbol{\mu}_{ij} = \left(\mu_{ij1}, \mu_{ij2}, ..., \mu_{ijK_{ij}} \right)^T$ is a $K_{ij} \times 1$ mean vector
- Σ_{ij} is a $K_{ij} \times K_{ij}$ variance-covariance matrix

for the $(i,j)^{th}$ unit.

If μ_{ijk} does not depend on time k, then the model simplifies to

$$\boldsymbol{\mu}_{ij} = \left(\mu_{ij}, \mu_{ij}, ..., \mu_{ij} \right)^T$$

and our interest centres on differences between the collection of μ_{ij}s.

We have encountered this situation previously when studying **ANOVA;** the distinction here is the correlation in the data.

- Classical ANOVA: **"BETWEEN SUBJECTS" -** data conditionally independent given group classification
- Here: **"WITHIN SUBJECTS" -** data correlated

The within-subjects design introduces **nuisance** parameters such as the parameters in Σ_{ij}; some simplifications, such as

$$\Sigma_{ij} = \Sigma_i \qquad \text{common covariance for group } i$$

or

$$\Sigma_{ij} = \Sigma \qquad \text{common covariance for all groups}$$

can be made (and tested).

7.2 WITHIN SUBJECTS ANOVA

The within subjects or **repeated measures** approach is used for several reasons:

- **Clinical:** some research hypotheses require repeated measures. Longitudinal research, for example, measures each sample member at each of several ages. In this case, age would be a repeated factor
- **Statistical:** in cases where there is a great deal of variation between sample members, error variance estimates from standard ANOVAs are large. Repeated measures of each sample member provides a way of accounting for this variance, thus reducing error variance.
- **Economic:** when sample members are difficult to recruit, repeated measures design are economical because each member is measured under all conditions.

Repeated measures ANOVA can also be used when sample members have been matched according to some important characteristic. Here, matched sets of sample members are generated, with each set having the same number of members and each member of a set being exposed to a different random level of a factor or set of factors. When sample members are matched, measurements across conditions are treated like repeated measures in a repeated measures ANOVA.

EXAMPLE: Suppose that a group of depressed subjects is selected, and their levels of depression measured. Suppose then that these subjects are arranged into pairs having similar depression levels; one subject from each matching pair is then given a treatment for depression, and afterwards the level of depression of the entire sample is measured again. ANOVA comparisons between the two groups for this final measure would be most efficient using a repeated measures ANOVA. In this case, each matched pair would be treated as a single sample member. As with any ANOVA, within subjects, or **repeated measures** ANOVA tests the equality of means.

NOTE: We should be clear about the difference between a **repeated measures** design and a **multivariate** design (see Chapter 8) .

- In the **repeated measures** design, each trial represents the measurement of the **same characteristic under a different condition**.
- In the **multivariate design**, each trial represents the measurement of a **different characteristic**. It is generally inappropriate to test for simple mean differences between measurements of different characteristics.

7.3 WORKED EXAMPLE

(taken from `http://www.utexas.edu/cc/docs/stat38.html`)

A health researcher wants to investigate the impact of dietary habits and types of exercise on individuals' **pulse rates** over time. To investigate these issues, a sample of individuals is collected, and grouped according to their **dietary preferences** D: (either meat eater or vegetarians) Each diet category is then split into three groups, with each group assigned one of three types of **exercise** E: (aerobic stair climbing, squash, and weight training). This, design has two **between-subjects** grouping factors: dietary preference and exercise type.

In addition to these between-subjects factors, a single **within-subjects** factor is to be included. Each subject's pulse rate will be measured at **three** exercise intervals (immediately after warm-up exercises, after jogging, and after running). Thus, **intensity of exertion** I is the within-subjects factor in this design. The order of these three measurements will be randomly assigned for each subject.

NOTE: all the factors described can be considered **fixed effects** (and not **random effects)**. The trials and groups were selected because of the research hypothesis. The levels of a random effect are chosen at random from a population of possible levels; random effects can not be appropriately analyzed with the method being described.

ID	DIET	EX.	WEIGHT	PULSE 1	PULSE 2	PULSE 3
1	Meat Eater	Stair	75.0	86	115	119
2	Meat Eater	Stair	85.6	72	117	129
3	Meat Eater	Stair	61.2	78	106	132
4	Meat Eater	Stair	72.2	68	108	129
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	Meat Eater	Stair	62.8	87	103	125
11	Meat Eater	Squash	74.7	90	127	128
12	Meat Eater	Squash	79.0	75	123	141
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	Meat Eater	Weights	70.1	61	185	204
31	Vegetarian	Stair	80.6	86	121	125
32	Vegetarian	Stair	79.7	68	105	143
⋮	⋮	⋮	⋮	⋮	⋮	⋮

There are four questions to be addressed:

- **Within-Subjects Main Effect**
  - Does exertion intensity influence pulse rate? (Does mean pulse rate change across the trials for exertion intensity?) This is the test for a within-subjects main effect of intensity.

- **Between-Subjects Main Effects**
  - Does dietary preference influence pulse rate? (Do vegetarians have different mean pulse rates than meat eaters?) This is the test for a between-subjects main effect of dietary preference.
  - Does exercise type influence pulse rate? (Are there differences in mean pulse rates between stair climbers, squash players, and weight trainers?) This is the test for a between-subjects main effect of exercise type.

- **Between-Subjects Interaction Effect**

  - Does the influence of exercise type on pulse rate depend on dietary preference? (Does the pattern of differences between mean pulse rates for exercise-type groups change for each dietary-preference group?) This is the test for a between-subjects interaction of exercise type by dietary preference. Note that other formulations of this interaction are equivalent. This hypothesis can also be expressed as "Does the influence of dietary preference depend on exercise type?".

The **interaction hypotheses** can be interpreted as follows; we may wish to test that vegetarian squash players have lower pulse rates than all meat eaters and other vegetarians; is something unique in the combination of a vegetarian diet and squash exercise that produces an unusually low mean pulse rate.

- **Within-Subjects by Between-Subjects Interaction Effects**
  - Does the influence of diet on pulse rate depend upon exertion intensity? (Does the pattern of differences between mean pulse rates for dietary-preference groups change at each intensity trial?) This is the test for a between-subjects by within-subjects interaction of dietary preference by exertion intensity. You might suspect, for example, that the mean pulse rate of meat eaters will increase more than the mean pulse rate of vegetarians as the intensity of exercise changes.
  - Does the influence of exercise type on pulse rate depend upon exertion intensity? (Does the pattern of differences between mean pulse rates for exercise-type groups change at each intensity trial?) This is the test for a between-subjects by within-subjects interaction of exercise type by exertion intensity.

- – Does the influence of dietary preference on pulse rate depend upon exercise type and exertion intensity? (Does the pattern of differences between mean pulse rates for dietary-preference groups change for some exercise-type group and for some intensity trial?) This is the test for a between-subjects by within-subjects interaction of dietary preference by exercise type by exertion intensity.

Each of these hypotheses relates to a different aspect of the model specification. Many of the statistical tests that are to be used are extensions of the ones used previously based on ANOVA considerations (in tests of equality of mean response), tests for equality of variance (Levene's Test etc.). The novelty here relates to testing hypotheses concerned with the correlation structure.

7.4 ASSESSING THE COVARIANCE STRUCTURE

Several methods are used to investigate the covariance structure in a repeated measures (or general multivariate) experiment.

7.4.1 BOX'S TEST OF EQUALITY OF COVARIANCE

Box's M-test tests for the equality of covariance matrices across G multivariate subgroups (defined by the fixed effects cross-categorization), that is

$$\begin{aligned} H_0 &: \Sigma_g = \Sigma \\ H_1 &: \Sigma_{g_1} \neq \Sigma_{g_2} \qquad \text{for some pair } (g_1, g_2) \text{ of sub-populations} \end{aligned}$$

against a general alternative.

The test statistic is M defined by

$$M = (n - G) \log |\mathbf{S}| - \sum_{g=1}^{G} (n_g - 1) \log |\mathbf{S}_g|$$

where

$$\mathbf{S} = \frac{\sum_{g=1}^{G} (n_g - 1) \mathbf{S}_g}{n - G} \qquad \mathbf{S}_g = \frac{\sum_{i=1}^{n_g} \left(\mathbf{y}_{gi} - \overline{\mathbf{y}}_g\right) \left(\mathbf{y}_{gi} - \overline{\mathbf{y}}_g\right)^T}{n_g - 1}$$

and where $\overline{\mathbf{y}}_g$ is the (vector) sample mean, $\mathbf{S}_g$ is the sample covariance matrix for subgroup g, and $\mathbf{S}$ is the pooled sample covariance matrix, and

$$n = n_1 + ... + n_g$$

is the total sample size. The null distribution for this test is the $Fisher - F$ distribution.

7.4.2 MULTIVARIATE TESTS

Multivariate tests are joint tests of the **significance of a main effects and within subjects effects**; the tests reported in SPSS are

- Pillai's Trace
- Wilks's Lambda
- Hotelling's Trace
- Roy's Largest Root

The details of these tests can be largely ignored; the important thing to note is that they follow the usual procedure of statistical hypothesis testing; a test statistic is derived (usually based on some transformation or eigen-representation of the sample covariance matrices) and the suprisingness of the observed test statistic is assessed against some null distribution. SPSS reports the appropriate $p-$values.

Which test is preferable ?

- Schatzoff (1966)
  - Roy's largest-latent root was the most sensitive when population centroids differed along a single dimension, but was otherwise least sensitive.
  - Under most conditions it was a toss-up between Wilks' and Hotelling's criteria.
- Olson (1976)
  - Pillai's criteria was the most robust to violations of assumptions concerning homogeneity of the covariance matrix.
  - Under diffuse noncentrality the ordering was Pillai, Wilks, Hotelling and Roy.

  - Under concentrated noncentrality the ordering is Roy, Hotelling, Wilks and Pillai.

- Best ?

  - When sample sizes are very large the Wilks, Hotelling and Pillai become asymptotically equivalent.

(http://www.gseis.ucla.edu/courses/ed231a1/notes3/manova.html)

7.4.3 MAUCHLY'S TEST OF SPHERICITY

The correlations between the different measurement times (not only successive but rather any time) are not usually the same, which prevents the use of the usual (Fisher F) test calculated as for ANOVA. The normal F test assumes the **sphericity** of the data, which means that variance of **all mutual differences of all possible pairs** of measuring times is the **same.** This is equivalent to an assumption that the $(i, j)^{th}$ element of the common covariance matrix for the random errors is some constant value σ

$$[\Sigma]_{ij} = [\Sigma]_{ji} = \sigma \qquad i \neq j$$

or that the correlation is some constant ρ.

Lack of sphericity causes concern about the ANOVA F-test; this is tested with **Mauchly's test**, whose null hypothesis is sphericity.. The rejection of sphericity does not prevent the analysis of variance, but the degrees of freedom should be adjusted before the ANOVA F-test result is reported.

The correction should be applied to the within-subject (time) effects and their corresponding error. Mean-Square values change, but the value of the F test does not, however the degrees of freedom values used to calculate the $p-$value do change, and can make a large difference to the conclusions made. If the number of measurement times is K, then the coefficient for correction of the degrees of freedom, ϵ, can take values between $1/(K-1)$ and 1, where 1 corresponds to the complete sphericity situation.

If the assumption of sphericity is strongly rejected, i.e. the coefficient of correction is near the lower limit $1/(K-1)$, we can use the **Greenhouse-Geisser** correction (rough limit $\epsilon < 0.75$). If the assumption of sphericity is broken just a little, i.e. ϵ is near one ($\epsilon > 0.75$), an adequate option is the **Huynh-Feldt** correction which is more liberal than the Greenhouse-Geisser correction (more sensitive to differences).

7.5 TESTS OF WITHIN-SUBJECTS EFFECTS

Univariate tests for the fixed effects (and contrasts) can also be carried out; this respects the true repeated measures aspect of the design, as opposed to the **multivariate tests** described above. **Repeated measures** ANOVA is used.

Repeated measures ANOVA carries the standard set of assumptions associated with an ordinary analysis of variance; extended to the matrix case: multivariate normality, homogeneity of covariance matrices, and independence. Repeated measures ANOVA is robust to violations of the first two assumptions. Violations of the independence assumption produce a non-normal distribution of the residuals, which results in invalid F ratios. The most common violations of independence occur when either random selection or random assignment is not used.

In addition to these assumptions, the univariate approach to tests of the within-subject effects requires the assumption of sphericity; if the sphericity assumption is not valid, conservative correction methods (such as Greenhouse-Geisser or Huynh-Feldt) should be utilized.

When sample sizes are small, the univariate approach can be more powerful, but this is true only when the assumption of a common spherical covariance matrix has been met.

Finally, a test of **homogeneity** of variances based on Levene's procedure should also be carried out.

7.6 MULTIVARIATE RESPONSES

A **multivariate response** experiment has much of the same structure as the repeated measures experiment that is described in the previous chapter. The principal extensions are that

- a number different variables can be measured for each experimental unit, possibly at different time points.
- simplifying assumptions necessary for repeated measures ANOVA (homogeneity, sphericity) can be relaxed; this may result in a less powerful analysis, but this is unavoidable if the simplyfing assumptions are not valid.