F Harrell

IX. ANALYSIS OF REPEATED MEASUREMENTS

In many experiments, multiple measurements are made on each independent experimental unit. When there are only two measurements per unit, the data may be analyzed by methods discussed previously by performing calculations on the difference or ratio of paired measurements. The statistical analysis is more complicated when more than two measurements are made. The multiple measurements on each unit are not independent and their degree of dependence or intercorrelation is usually unknown; it is difficult to determine how much weight each repeated measurement should be given in the overall analysis.

Repeated measurements actually represent multivariate observations from some high dimension statistical distribution, and at face value should be analyzed using a complicated multivariate analysis. However, in many situations the problem can be reduced to a simple one capable of being analyzed with methods already presented.

Examples of repeated measurement studies include experiments in which measurements are made on each subject at several time intervals or when each subject is studied under several condition.

Traditional Analyses of Repeated Measurements

Many investigators have analyzed repeated measurements using analysis of variance techniques. These methods are not appropriate. An an example, consider the analysis of ventilation volume of 8 subjects under 6 different temperatures of inspired dry air appearing in Deal, et al (1) in the table of data below. The authors performed a one-way analysis of variance to test whether there were any differences in ventilation volumes at different temperatures. For each subject, the order of temperatures was randomized so that fatigue or learning would not confound the study of temperature conditions.

Figure 1

From Table I of Deal et al (1979): Role of respiratory heat exchange in production of exercise-induced asthma. J Appl Physiol 46:467-475.

MINUTE VENTILATION VS TEMPERATURE-DRY GAS EXPERIMENTS

Ventilations in 1/min

Temp:	-10 °C	25	37	50	65	80	Mean	SD	Slope*	Signed-Rank
Subject										
1	74.5	81.5	83.6	68.6	73.1	79.4	76.8	5.7	- .9	-2
2	75.5	84.6	70.6	87.3	73.0	75.0	77.7	6.7	-2.0	-4
3	68.9	71.6	55.9	61.9	60.5	61.8	63.4	5.8	-10.4	-7
4	57.0	61.3	54.1	59.2	56.6	58.8	57.8	2.5	+ .4	1
5	78.3	84.9	64.0	62.2	60.1	78.7	71.4	10.5	-12.0	-8
6	54.0	62.8	63.0	58.0	56.0	51.5	57.6	4.7	-3.8	-5
7	72.5	68.3	67.8	71.5	65.0	67.7	68.8	2.8	-5.7	-6
8	80.8	89.9	83.2	83.0	85.7	79.6	83.7	3.7	-1.3	-3
Mean	70.2	75.6	67.8	69.0	66.3	69.1			-4.5	-34 sum
SD	9.8	11.0	11.1	11.0	10.3	10.8			4.6	204 sum of squares

* Slope of subject's ventilation volume on temperature x100

$$\text{Slope} = \sum_{i=1}^{6}(T_i-\bar{T})V_i \Big/ \sum_{i=1}^{6}(T_i-\bar{T})^2$$

$$\bar{T} = 41.167 \quad \sum(T_i-\bar{T})^2 = 7071.167$$

$$\text{Slope} \times 100 = -.0724\,V_1 - .0229\,V_2 - .0059\,V_3 + .1249\,V_4 + .3370\,V_5 + .5492\,V_6$$

Author's test for effect of temperature:
One-way analysis of variance: $F_{5,42} = .72$, $p>.5$

Using slopes to test for trends:
Wilcoxon signed rank test: $u=\bar{S}/(n+1)=-.47$, $p=.016$
One-sample t-test: $t_7 = -2.77$, $p=.028$

The F statistic from the analysis of variance with 5 and 42 degrees of freedom was .72 and p>.5. The authors concluded that no differences existed between temperatures. The flaws in this analysis are:

1. The observations within a single subject are not independent. Also, the normality of the distributions of ventilation volumes has not been checked.
2. The analysis of variance F statistic makes use of only the following information: the column means, column standard deviations, and sample sizes. The subject identifications are not used. Within each column, the subjects' measurements could be randomly rearranged without affecting F. The analysis does not remove variation among subjects by considering each subject as his own control, making use of the fact that variation within subjects is usually less than the variation between subjects. This is indeed the case as can be seen by comparing the row standard deviations to the column standard deviations. The standard deviations of measurements within subjects across temperature is less than the variation within temperatures across subjects.
3. The columns are written in order of increasing temperature. However, if the column headings were randomly reassigned, F would be unchanged. It would be reasonable to expect beforehand that there would be a consistent relationship between temperature and ventilation. The analysis of variance loses information about the temperature continuum. Also, the original hypothesis is not focused. The test for all possible differences has 5 degrees of freedom. It is not concentrated enough to pick up trends with temperature.

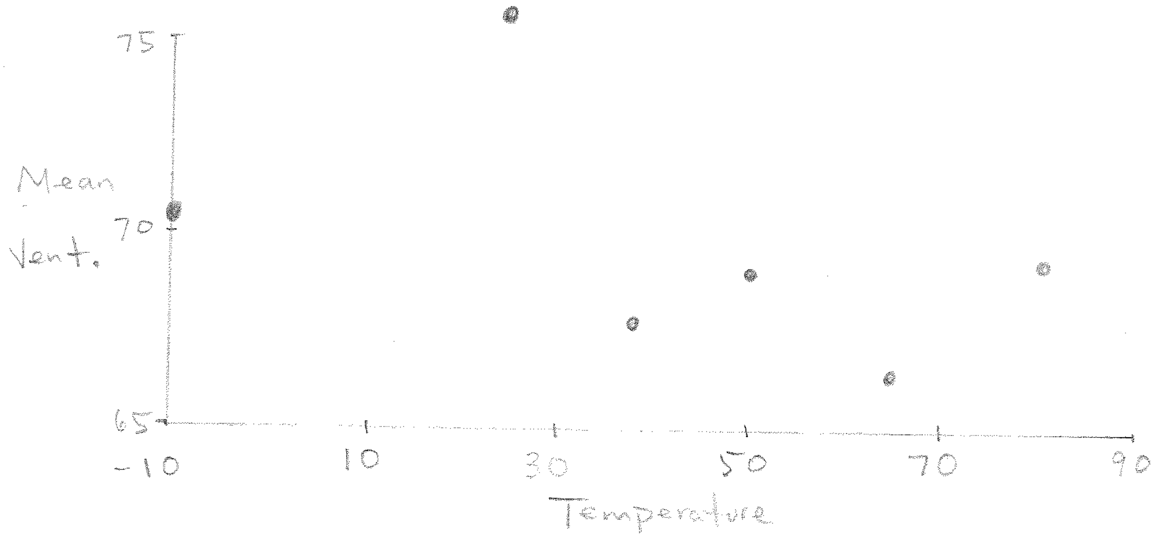
In summary, the F test makes poor use of available information in the data and makes unverifiable assumptions.

An analysis called repeated measures analysis of variance or two-way analysis of variance solves problem #2. It treats each subject as her own control by subtracting each subject's overall mean measurement from each of her measurements. However, problems 1 and 3 remain and the repeated measures analysis makes additional restrictive and unverifiable assumptions, namely that the array of measurements for each subject follow a multivariate normal distribution <u>and</u> all pairs of measurements are equally correlated.

Solutions

Many studies such as the one just described fall under an area called "growth curve analysis", which is virtually the same as the analysis of dose-response and time-response relationships. Ghosh, Grizzle, and Sen (2) provide a general method of analysis for these situations. We will now analyze the ventilation data using their method. This method solves problem #3 above by testing for some overall trend of ventilation with temperature and testing a hypothesis having one degree of freedom. Problem #1 is solved by computing the measure of trend separately for each subject and then analyzing the 8 trend measurements, which by their construction are independent.

The first step in analysis is to graph the data. Below the mean ventilation at each temperature is plotted.



This is only a crude display as the variation between subjects can sometimes obscure the trend. It would be better to graph each subject's curve or to subtract each subject's mean response before computing the mean within a temperature. From the graph above we see some tendency for ventilations to decrease with increasing temperature. Let us suppose that the tendency can be satisfactorily described using a linear relationship.

The next step in the analysis is to compute the slope of temperature vs. ventilation separately for each subject. Denote the six temperatures by T_1, T_2, $T_3, \ldots,$ T_6 and for a given subject denote the six corresponding ventilation volumes by V_1, $V_2, \ldots,$ V_6. Recall that the formula for the least squares slope is

$$\sum_{i=1}^{6} (T_i - \overline{T}) V_i / \sum_{i=1}^{6} (T_i - \overline{T})^2.$$

which can be written $\sum_{i=1}^{6} w_i V_i$ where the coefficient of V_i is

$$w_i = (T_i - \overline{T}) / \sum_{i=1}^{6} (T_i - \overline{T})^2.$$

Here $\overline{T} = 41.167$ and $\Sigma(T_i - \overline{T})^2 = 7071.167$. Since the temperature points are identical for each subject, the computations can be minimized by computing the weights w_i once. The resulting equation is

$$\text{slope X 100} = -.0724\ V_1 - .0229\ V_2 - .0059\ V_3 + .1249\ V_4 + .3370\ V_5 + .5492\ V_6.$$

If one were only interested in the statistical test for trend and not in quantifying the trend, one could ignore the denominator of w_i and weight each Y-axis point with a weight equal to the difference between the corresponding X-axis point and the mean X-axis point. Alternatively, the slopes could be computed separately for each subject using a computer program. This approach is more flexible and it doesn't require the temperature settings to be exactly the same for each subject.

The slopes (X 100) are displayed on figure 1. The mean slope X 100 is 4.6. Note that the intercepts are being ignored; the intercepts absorb the varying subject baselines. Assuming the slopes to be normally distributed, one could test for the population mean slope being zero by the one-sample t-test. The t-statistic is $-4.5/(4.6/\sqrt{8}) = -2.77$ with 7 d.f. ($p=.028$). One could instead perform the Wilcoxon signed rank test assuming only that the slopes are symmetrically distributed about their median. The mean signed rank is $S=34$ and the index of tendency of slopes to be unequally distributed about zero is $W = \overline{S}/(n+1) = -.47$ (the minimum possible value of W is $-.5$). The approximate normal test statistic is $-34/\sqrt{204} = -2.38$ with $p=.017$. The exact p-value is .016.

Thus there is evidence that the median slope if not zero, i.e. there is a tendency for the minute ventilations to decrease with temperature. Since the average slope is -.045, a point estimate of the amount of decrease in ventilation volume per 25°C increase in air temperature is 25 x (.045) = 1.1 l/min. There is strong evidence that the slope is non-zero; the physiologic importance of this relationship has to be assessed by the investigator.

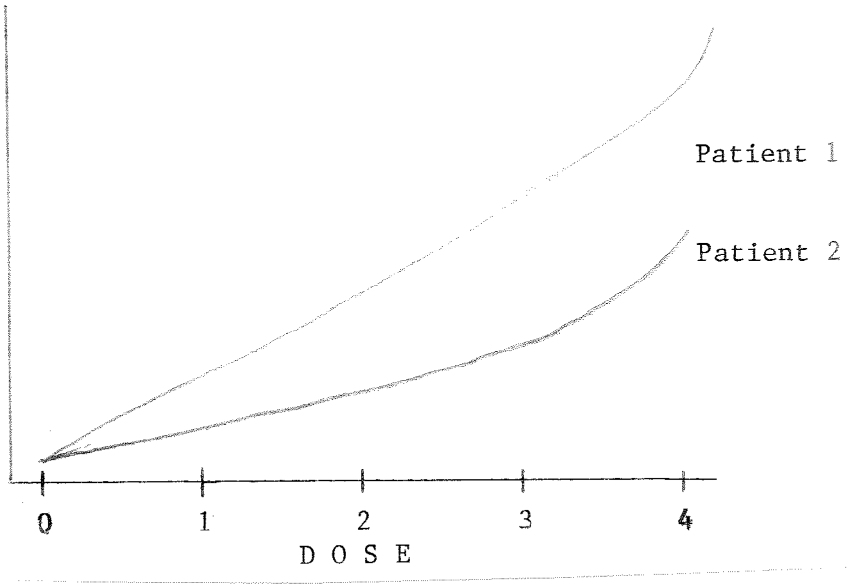
An alternate analysis could be performed that doesn't assume the relationship between temperature and ventilation to be linear, only monotonic. This analysis is efficient if there are at least 5 measurements per subject, and is done by computing the Kendall t_a or Spearman ρ rank correlation between temperature and ventilation separately for each subject and testing these correlations with the Wilcoxon sign rank test. In the table below the ranks of ventilations across increasing temperatures are displayed for each subject with the resulting value of ρ.

Subject	Ranks	ρ
1	3 5 6 1 2 4	-.257
2	4 5 1 6 2 3	-.257
3	5 6 1 4 2 3	-.543
4	3 6 1 5 2 4	-.086
5	4 6 3 2 1 5	-.314
6	2 5 6 4 3 1	-.371
7	6 4 3 5 1 2	-.771
8	2 6 4 3 5 1	-.257

Since each subject's ρ is negative, the exact 2-sided p-value is 2^{-7} = 1/128 = .0078 using the Wilcoxon signed rank test. Note that the rank correlation method is not sensitive for testing for variation in slopes when individual correlations are likely to be high. In that case, all rank correlations will be unity and tests for group differences will not detect anything. However, tests for significant correlations will be sensitive in this case.

Other Examples

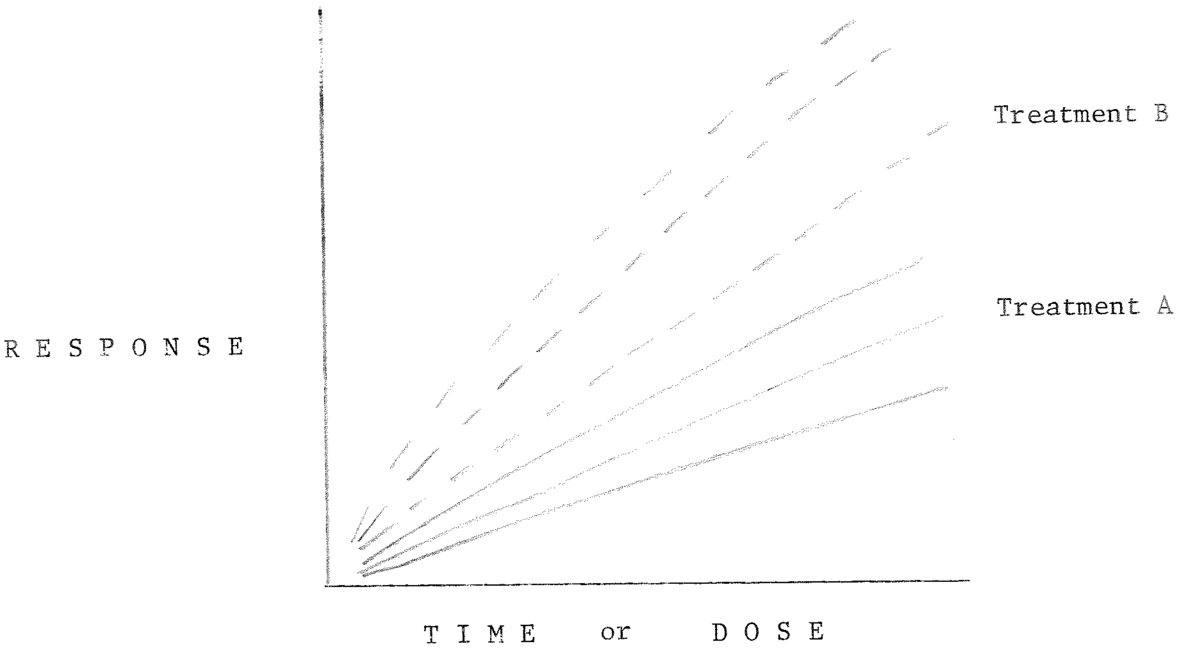
Suppose that each patient is given a series of dosages of a drug. The investigator may wish to test for a dose-response relationship such as the one shown in the next figure.



Dose-response relationships are often examined by testing for example dose 1 vs. dose 0, dose 2 vs. dose 0, etc, using a paired test. Researchers will often declare at which dose the response is significantly different from the zero dose response. Such multiple tests should be avoided. It is better to test for an overall dose-response trend by computing a single measure of response for each patient and testing these response measures.

To describe the effective dose, one could compute for each patient the dose at which the response exceeds a certain level. The median threshold dose and perhaps the range or inter-quartile range would nicely describe the distribution of effective doses across patients.

For another example, suppose that there are two groups of patients each given a separate treatment. For each treatment, each patient may receive various doses of a drug or each may be studied at various times under identical treatment conditions.

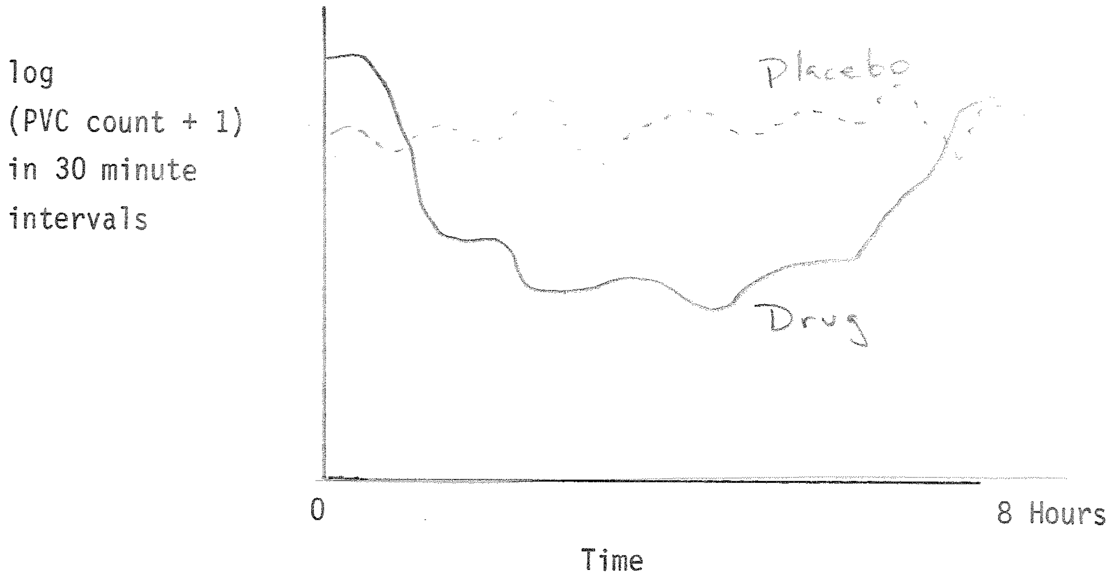


The difference between treatments can be assessed by for example computing a slope for each patient and comparing the treatment A slopes with the treatment B slopes using the Wilcoxon 2-sample test.

Note that as always, evidence for a treatment difference is based on the consistency of response patterns across experimental units (i.e. patients). If this were not the case, one might as well study one animal at dozens of times or doses.

A More Complicated Example

A new drug is designed to reduce the number of premature ventricular contractions (PVC's) of the heart for patients suffering from significant arrhythmia disorders. To test the hypothesis of a lowering of PVC's by the drug, each of 8 patients received a dose of the drug or of a placebo and was studied for an 8-hour period. One week later, the patient received the alternate treatment and a repeat study was performed. The order of treatments was randomized for each patient to minimize time effects (this experimental design is called a crossover design). Previous studies have demonstrated that logarithms of (PVC count +1) have a well-behaved distribution. A typical graph of the responses from one patient is given below.



The drug time response is nonlinear. Without assuming anything about the shape of the response curve, a useful way to quantify the treatment response is to compute the area under each curve. Because of marked day to day variability in PVC's within patients it is desirable to use the time zero count as a control for each day. We can incorporate this control period by computing the area after shifting the curve vertically so that the time zero log (PVC + 1) is zero. Each patient was measured at each 30 minute period for both days, so this area is proportional to the average log (PVC +1) at time > 0 minus log (PVC + 1) at time=0. These areas are shown below.

	AREA		
Patient	Placebo	Drug	Difference
1	-1.6	-36.7	35.0
2	2.9	-2.3	5.3
3	-.8	-10.7	9.9
4	-1.2	-17.3	16.1
5	-7.2	-23.0	15.8
6	1.7	-34.9	36.6
7	4.7	1.2	3.4
8	1.8	-36.2	38.0

The areas are paired since each patient underwent two treatments. The differences in areas measure the overall drug response. The exact 2-tailed p-value using the Wilcoxon signed rank test is 2^{-7} = .0078 indicating a significant drug effect in reducing PVC's.

Had some patients not had PVCs measured for each time period, areas cannot be computed by this method. Another way to estimate areas is to use the trapezoidal rule (see Appendix) or to fit a function to each patient's curve using least squares. These methods do not require the time points to be the same for each patient although some caution should be taken when trying to extrapolate estimates for patients studied a very short time.

Quite often a quadratic equation adequately fits the response curve. The equation for a given patient is $\log(PVC + 1) = at^2 + bt + c$, where t denotes time and a, b, and c denote least squares estimates. In order to use the time zero response as a control point and to calculate all other estimated responses with reference to that baseline, we ignore the intercept c. The area from 0 to T under the baseline point then is $aT^3/3 + bT^2/2$. The area is calculated separately for each study day for each patient with T set to a reasonable constant (here T=8 hours). The analysis of the treatment effect proceeds as before now that areas have been calculated that effectively use interpolation to estimate responses at time points for which measurements were not made.

Other Example Problems

Other repeated measurement problems are described below, along with sketches of their solutions.

Example Problem 1

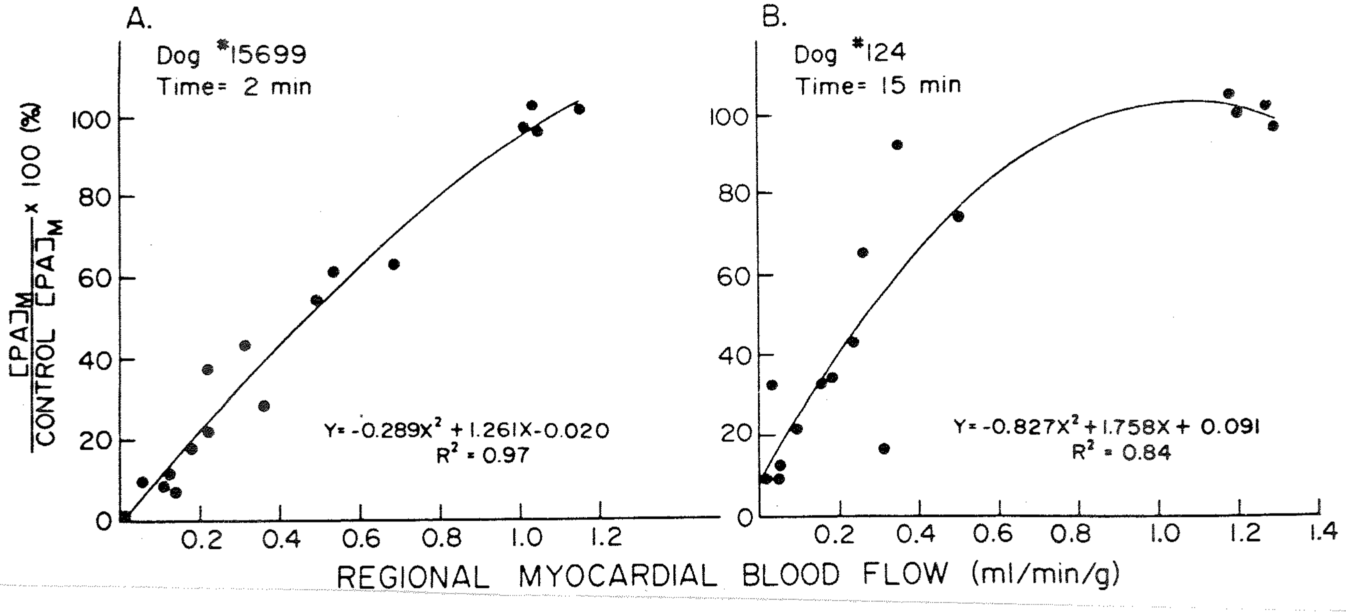
An experiment was conducted to assess the effect of the amount of intravenous injection of a drug on patients' heart rates measured over time. Each of the patients was given only one dose of the drug (either 0, 1, 2, or 3 µg/Kg) and had heart rate measured at 11 times varying from 0 to 35 minutes after injection.

Each patient's baseline (time 0) measurement was subtracted from the remaining 10 measurements to adjust for baseline differences. The primary hypothesis of interest was that the time until return to the baseline state increased with increasing dose. To test this hypothesis, a linear regression equation was fitted to each patients 10 time-heart rate pairs and the x-intercept, the negative of the intercept estimate described by the slope estimate, was calculated. The Kendall rank correlation test was used to test whether the x-intercepts increased with dose.

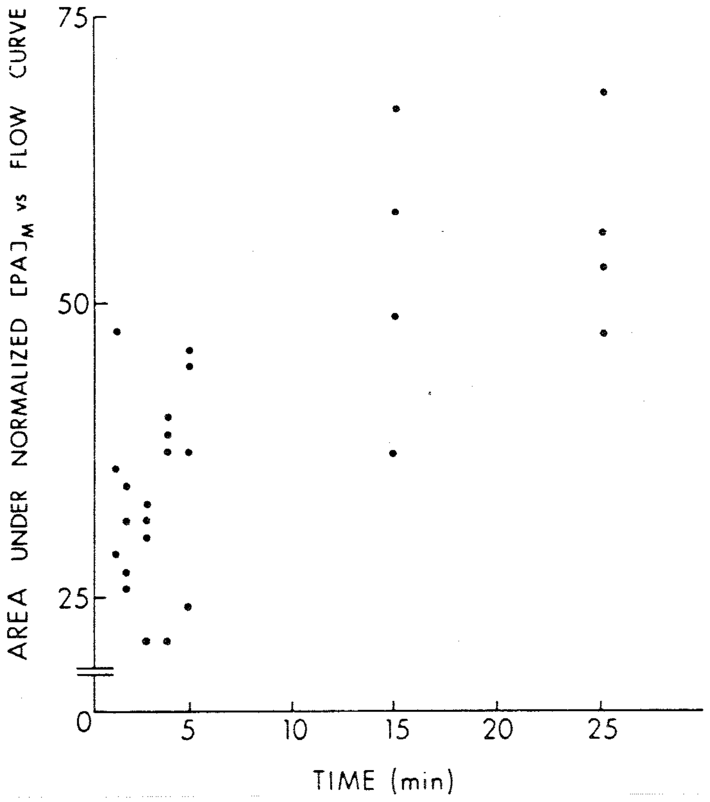
In a secondary analysis, the peak heart rate after completion of infusion (after 5 minutes), normalized for the baseline measurement, was computed. This peak response was then correlated with dose.

Example Problem 2

In a study (3) to delineate the time course of procainamide accumulation in myocardium and its relationship to regional myocardial blood flow, 28 dogs were sacrificed at different times (1.5-15 minutes) after onset of drug infusion. Regional blood flow and procainamide concentration were measured in myocardium sections. As part of the study it was of interest to analyze how the relationship between blood flow and procainamide concentration in the myocardium sections was influenced by the length of time from infusion of the drug. An average of 16 sections were obtained from each dog, with an adequate range of blood flows. Data from two dogs are shown in the following figures.



The area under the flow-concentration curve from 0 to .8 ml/min/g was used as a measure of the overall relationship. To estimate the area, a quadratic regression equation of the form $C=aF^2 + bF + c$ was fitted separately for each dog. The estimate of area is then $A = aF^3/3 + bF^2/2 + cF$, where .8 ml/min/g is substituted for F. [The trapezoidal rule could have been used to calculate the area although differing flow ranges might pose a problem.] The relationship between area and time is shown in the following graph.



The normal scores rank correlation test (similar to the Spearman correlation test) was used to test whether the area increased with time since drug administration, yielding $p<.01$. The inference was that the area under the curve increased with time, and that the longer the time following drug administration, the higher was the normalized myocardial procainamide concentration at any given blood flow.

Example Problem 3

An *in vitro* electrophysiologic study (4) was undertaken to determine the mechanism of action of disobutamide. Part of the experiment involved studying the effects of disobutamide at three doses (0 (control), 0.3 and 3.0 or 1.0 and 10.0 μg/ml)) on action potential characteristics of isolated Purkinje fibers, using intracellular microelectrodes. Fifty-two dogs were sacrificed and a total of 16 Purkinje fibers were obtained from their hearts. When two fibers could be obtained, the dose sequences 0.3 and 3 μg/ml and 1.0 and 10.0 μg/ml were both used, with the selection of fibers for the two dose sequences randomized. Different fibers, even from the same dog, were considered to be independent experimental units. Sixteen fibers were used in the study.

One of the action potential characteristics analyzed was Vmax, the maximum rate of rise of the transmembrane potential in phase 0. Vmax was recorded at cycle lengths (CL) of 1000, 500, 400, and 320 msec for each dose of disobutamide for each Purkinje fiber. Thus for each fiber 12 measurements were made (3 doses x 4 CLs). Graphs revealed that the relationships of Vmax to dose and to the reciprocal of CL were approximately linear.

The major hypothesis tested was that of no drug effect on Vmax. Since each fiber was exposed to the same 4 CLs, the Vmax values for a given dose were averaged over the 4 measurements corresponding to 4 CLs. Let Y_i be the mean response over 4 CLs for dose level i, with i=1 corresponding to dose 0 (control), i=2 corresponding to dose 0.3 or 1 μg/ml, and i=3 corresponding to dose 3.0 or 10.0 μg/ml. The drug effect was tested by testing for a linear trend of Y_1, Y_2, Y_3 vs. dose, separately for each dose sequence. For either sequence (since doses are of the form 0, d, 10d for each), the slope of Vmax vs. dose is proportional to $S = -11Y_1 - 8Y_2 + 19Y_3$. For each dose sequence, S was calculated for each of 8 fibers. The drug effect (i.e., the effect of increasing dose) can be tested by testing whether S has median 0, using the

Wilcoxon signed rank test. A one-sample t-test was actually used, and a significant drug effect was found for both dose ranges.

A second analysis was undertaken to test whether the effect of drug depended on CL, i.e., whether there was a dose x CL interaction. Separately for each of 8 fibers, a multiple linear regression model of the form $Y = a + b_1$ dose $+ b_2 \times CL^{-1} + b_{12} \times$ dose $\times CL^{-1}$ was fitted to the 12 dose - CL combinations. Thus 8 b_{12} values were obtained. Ho: median $b_{12} = 0$ can be tested with the Wilcoxon signed rank test. A one-sample t-test was actually used, and no evidence was found for a CL - dependent change in Vmax with increasing doses.

Example Problem 4

An investigator interested in a diurnal response of a circulating hormone measured serum hormone levels in 20 guinea pigs at 6AM, 12PM, 6PM, and 12AM. Guinea pigs ranged in age from 1 month to 1 year. The investigator hypothesized that the peak hormone level occurred later in the day as animals aged, and after a certain age the animals were more likely to have a biphasic response.

An ordinal response variable was constructed which essentially captured the time until the peak response for each animal, coding a biphasic response as the highest response level. Specifically,

$Y=0$ if peak occurred at 6 AM,

$Y=1$ if peak occurred at 12 PM,

$Y=2$ if peak occurred at 6 PM,

$Y=3$ if peak occurred at 12 AM,

$Y=4$ if the two highest values: (a) different by less than 1.5 μg/100ml, (b) occurred at least 12 hours apart, and (c) averaged at least 3 μg/100ml higher than the average of the two lowest values.

An ordinal logistic model was used to test the hypothesis that the age of guinea pigs is associated with increasing Y.

Summary

In analyzing repeated measurement studies, the following recommendations are made:

1. Compute a reasonable measure of trend (slope, area, t_a, ρ, etc.) for each curve.

2. Analyze these summary measures as if they were original data. Use the fact that summary measures computed from different experimental units are independent.
3. Test trends with a rank test.

The major assumption made by this procedure is that trend information is adequately captured by the summary statistics. No assumption is made concerning the correlation structure within an experimental unit. No hard distributional assumptions are made.

Appendix

Calculating the Area Under a "Curve" Without Assuming a Form for the Curve

Suppose that n X-Y pairs are measured: (X_1,Y_1), (X_2,Y_2), ..., (X_n, Y_n), listed in order of increasing X values. By connecting successive Y points with straight lines, one can obtain an estimate of the area under the X-Y curve using the trapezoidal rule. The area A is given by

$$A = .5\ [(X_2-X_1)Y_1 + (X_3-X_1)Y_2 + (X_4-X_2)Y_3 + \ldots + (X_n-X_{n-2})Y_{n-1} + (X_n-X_{n-1})Y_n].$$

When the spacing of successive Xs is a constant w,

$$A = .5w\ [Y_1+2\ Y_2 + 2\ Y_3 + \ldots + 2\ Y_{n-1} + Y_n].$$

For the case when n is odd and there is a common spacing w, a better estimate can be obtained by fitting piecewise quadratic curves to the X-Y pairs. This method, called Simpson's rule, yields

$$A = (2w/3)\ [Y_1 + 4\ Y_2 + 2\ Y_3 + 4\ Y_4 + \ldots + 2\ Y_{n-2} + 4\ Y_{n-1} + Y_n].$$

If for different experimental units the span of X, X_n-X_1, differs, one may normalize for this difference by dividing A by X_n-X_1 to obtain the mean Y value (considering Y as a smooth function of X). Such normalization is risky if the spans vary greatly.

Problem

In a study conducted to test whether pH alters action potential characteristics when patients are given a drug, each of 25 patients had Vmax measured at up to four pH levels:

Patient #	6.5	6.9	7.4	7.9
1		284	310	326
2			261	292
3		213	224	240
4		222	235	247
5			270	286
6			210	218
7		216	234	237
8		236	273	283
9	220	249	270	281
10	166	218	244	
11	227	258	282	286
12	216		284	
13			257	284
14	204	234	268	
15			258	267
16		193	224	235
17	185	222	252	263
18		238	301	300
19		198	240	
20		235	255	
21		216	238	
22		197	212	219
23		234	238	
24			295	281
25			261	272

Test the null hypothesis that there is no relationship between pH and Vmax.

References

1. Deal EC, McFadden ER, Ingram RH, Strauss RH, Jaeger JJ: Role of respiratory heat exchange in production of exercise-induced asthma. J Appl Physiol 46:467-75, 1979.
2. Ghosh M, Grizzle JE, Sen PK: Nonparametric methods in longitudinal studies. J Am Statist Assoc 68:29-36, 1973.
3. Wenger TL, Browning DJ, Masterton CE, Abou-Donia MB, Harrell FE, Bache RJ, Strauss HC: Procainamide delivery to ischemic canine myocardium following rapid intravenous administration. Circ Res 46:789-95, 1980.
4. Dohrmann ML, Harrell FE, Strauss HC: The effects of disobutamide on electrophysiologic properties of canine cardiac Prukinje fibers and papillary muscle. J Pharmacol Exp Ther 217:549-54, 1981.