1 Introduction and Motivation

1.1 Objective of this course

OBJECTIVE: The goal of this course is to provide an overview of statistical models and methods for the analysis of ***longitudinal data***; that is, data in the form of ***repeated measurements*** over time or other factor on each ***individual*** or ***unit*** (human subject, plant, plot, sample, etc.) in a sample from a population of interest.

Data are collected routinely in this fashion in a broad range of applications, including agriculture and the life sciences, health sciences research, and physical science and engineering. For example

- In agriculture, a measure of growth is taken on experimental plots ***weekly*** over the growing season. Plots are assigned to be treated with different fertilizers at the start of the season.
- In a study of human immunodeficiency virus (HIV) infection, a measure of viral load (roughly, concentration of HIV present in the blood) is made ***monthly*** on infected patients. Patients are assigned to take different "cocktails" of antiretroviral treatments at the start of the study.

A defining characteristic of these examples is that the ***same*** response is measured ***repeatedly*** on each unit; e.g., viral load is measured repeatedly on the same subject. This particular type of data structure is the focus of this course.

The scientific questions of interest often involve not only the usual kinds of questions, such as how the ***mean response*** differs across treatments, but also how the ***change in mean response over time*** differs and other features of the relationship between response and time. Thus, it is necessary to represent the situation in terms of a ***statistical model*** that acknowledges the way in which the data were collected to address these questions. Complementing the models, specialized methods of analysis are required.

Because the study of change is fundamental across almost all scientific disciplines, studies in which longitudinal data are collected have become ubiquitous, and interest in the most appropriate ways to represent and interpret longitudinal data has grown tremendously. In this course, we will study approaches to modeling these data, and we will explore the associated classical and more modern approaches to analyzing them in detail.

TERMINOLOGY: Although the term ***longitudinal*** suggests that data are collected over ***time***, the models and methods we will discuss are more broadly applicable to any kind of ***repeated measurement*** data. That is, although repeated measurement most often takes place over time, this is not the only way that measurements can be taken repeatedly on the same individual or unit. For example,

- Individuals may be human subjects. On each subject, ***prothrombin time***, a measure of how long it takes blood to clot, is measured on several occasions, each involving administration of a different dose of an anti-coagulant agent. Thus, a subject is measured repeatedly over ***dose***.
- Units may be trees in a forest. For each tree, measurements of the diameter of the tree are made at several different points along the trunk of the tree. Thus, the tree is measured repeatedly over ***positions*** along the trunk.
- Individuals may be pregnant female rats. Each gives birth to a litter of pups, and the birthweight of each pup is recorded. Thus, the rat is measured repeatedly over each of her ***pups***.

The third example differs from the other two in that there is no natural ***order*** to the repeated measurements.

Thus, the methods apply more broadly than the strict definition of the term ***longitudinal data*** indicates – this term will mean, to us, data in the form of ***repeated measurements*** that might be over time, but might alternatively be over some other set of conditions. Because time is most often the condition of measurement, however, many of our examples will indeed involve repeated measurement over time, and we will often use the word ***time*** to refer generically to the repeated measurement condition.

We use the terms ***response*** or ***outcome*** to denote the repeated measurement or outcome of interest. Because longitudinal studies are frequently conducted with human or animal subjects, we use the terms ***unit***, ***individual***, and ***subject*** interchangeably.

1.2 Examples

To set the stage, we consider several data sets from a variety of applications. These not only provide concrete examples of longitudinal data situations, but serve to illustrate the range of ways that data are collected and the types of responses and questions that may be of interest.

EXAMPLE 1: The orthodontic study data of Potthoff and Roy (1964). This is a world famous data set that is used to introduce features of longitudinal data modeling and analysis. A study was conducted involving 27 children, 16 boys and 11 girls. On each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was made at ages 8, 10, 12, and 14 years of age. The pterygomaxillary fissure is a vertical opening in the human skull, depicted in Figure 1.1.

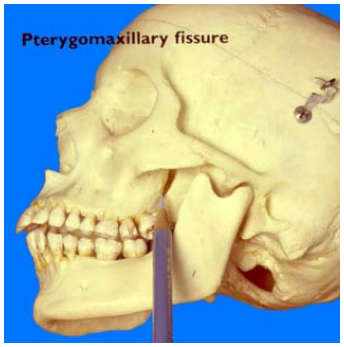

Figure 1.1: *Pterygomaxillary fissure.*

In Figure 1.2, the distance measurements are plotted against age for each child.

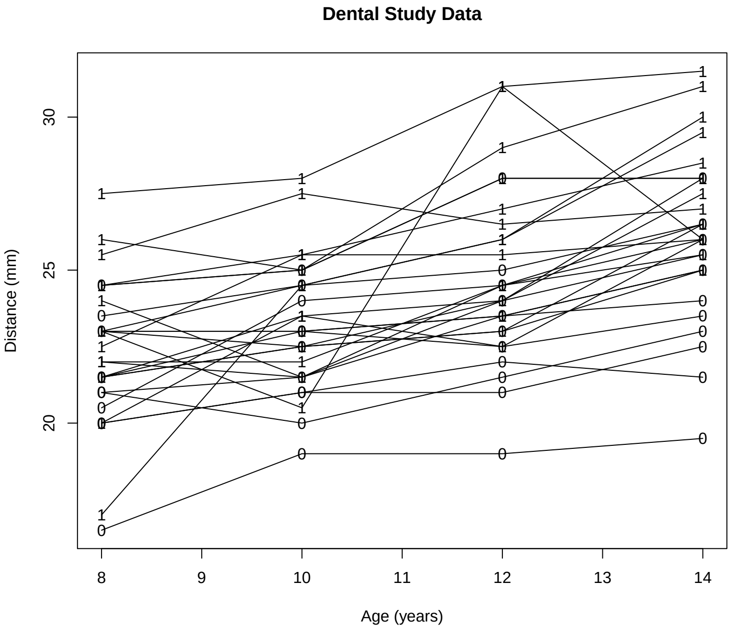

Figure 1.2: *Orthodontic distance measurements (mm) for 27 children at ages 8, 10, 12, 14. The plotting symbols are 0's for girls, 1's for boys.*

The plotting symbols denote girls (0) and boys (1), and the trajectory for each child is connected by a solid line so that individual child patterns can be seen. Plots like Figure 1.2 are often called ***spaghetti plots***, for obvious reasons.

The objectives of the study were to

- Determine if distances over time are larger on average for boys than for girls
- Determine if the ***rate of change*** of distance over time is different for boys and girls.

Several features are notable from the plot of the data:

- Each child has his/her own ***trajectory*** of distance as a function of age. For any given child, the trajectory looks roughly like a ***straight line*** , with some fluctuations. But from child to child, features of the trajectory (e.g., its ***steepness***), vary. Thus, the trajectories are all of similar form, but vary in their specific characteristics among children. Note the one unusual boy whose pattern fluctuates more profoundly than those of the other children and the one girl who is much "lower" than the others.
- The ***overall trend*** is for the distance measurement to ***increase*** with age. The trajectories for some children exhibit strict increase with age, while others show some intermittent decreases, but still with an overall increasing trend across the entire 6 year period.
- The distance trajectories for boys seem for the most part to be "***higher*** " than those for girls – most of the boy profiles involve larger distance measurements than those for girls. However, this is not uniformly true: some girls have larger distance measurements than boys at some of the ages.
- Although boys seems to have larger distance measurements, the ***rate of change*** of the measurements with increasing age seems similar. More precisely, the ***slope*** of the increasing (approximate straight-line) relationship with age seems roughly similar for boys and girls. However, for any ***individual*** boy or girl, the rate of change (slope) might be steeper or shallower than the apparent ***typical*** rate of change.

The foregoing observations are informal visual impressions. To address the questions of interest, it is clear that some formal way of representing these features is needed. Within such a representation, a formal way of stating the questions is required.

EXAMPLE 2: Vitamin E diet supplement and growth of guinea pigs. These data are reported by Crowder and Hand (1990, p. 27) and are from a study of the effect of a vitamin E diet supplement on the growth of guinea pigs. Fifteen guinea pigs were given a growth-inhibiting substance at the beginning of week 1 of the study (time 0, prior to the first measurement), and body weight was measured at the ends of weeks 1, 3, and 4. At the beginning of week 5, the pigs were randomized into 3 groups of 5, and vitamin E therapy was started. One group received zero dose of vitamin E, another received a low dose, and the third received a high dose. The body weight (g) of each guinea pig was measured at the end of weeks 5, 6, and 7.

In Figure 1.3, the data for the three dose groups are shown in spaghetti plots for each group; the plotting symbol is ID number (1–15) for each guinea pig.

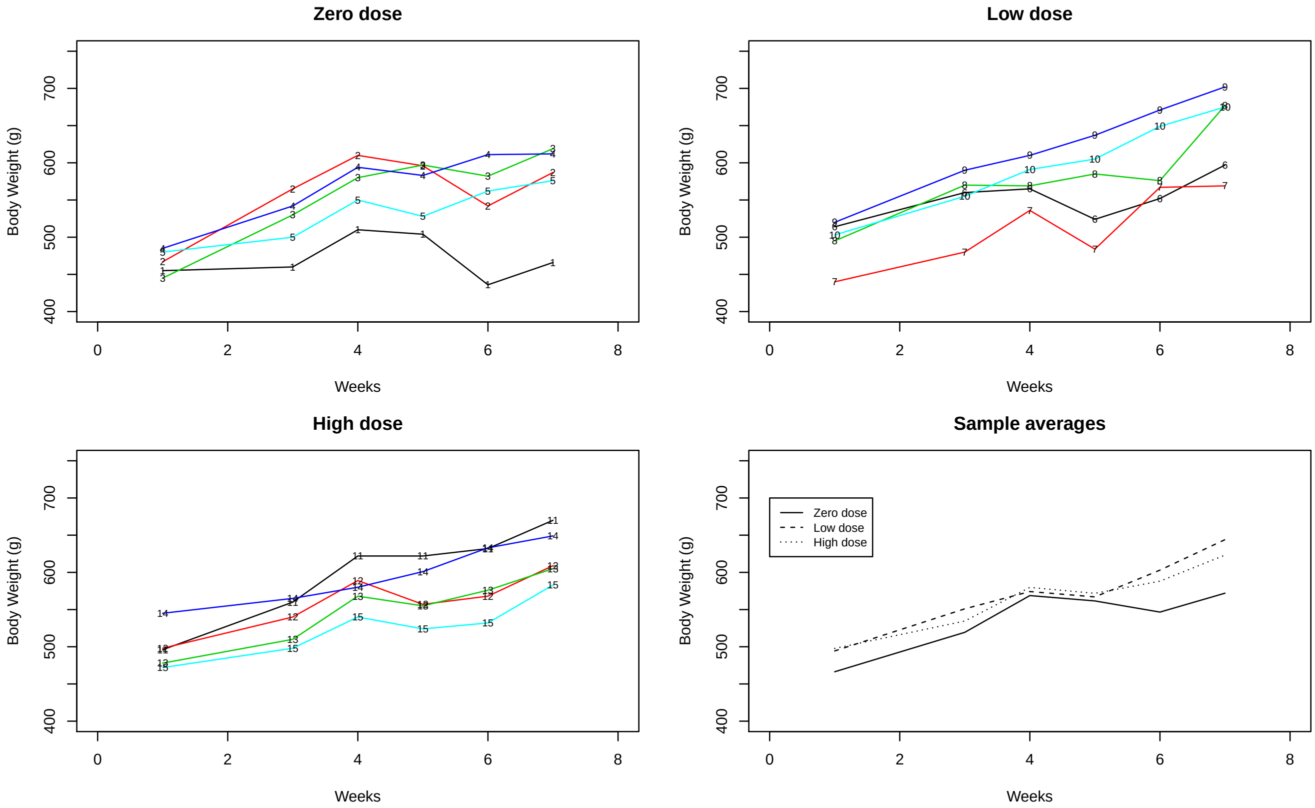

Figure 1.3: *Growth of guinea pigs receiving different doses of vitamin E diet supplement. Pigs 1–5 received zero dose, pigs 6–10 received low dose, pigs 11–15 received high dose.*

The primary objective of the study was to

- Determine if the ***growth patterns*** differ among the three groups.

As with the dental data, several features are evident:

- For the most part, the trajectories for individual guinea pigs seem to ***increase overall*** over the study period (although note pig 1 in the zero dose group). Different guinea pigs in the same dose group have different trajectories, some of which look like a straight line and others of which seem to have a "dip" at the beginning of week 5, the time at which vitamin E was added in the low and high dose groups.
- Trajectories for the zero dose group seem somewhat "***lower*** " than those in the other groups.
- It is unclear whether or not the rate of change in body weight on average is similar or different across dose groups. In fact, it is not clear that the pattern for either individual pigs or "on average" is a ***straight line***, so the rate of change might not be constant. Because vitamin E therapy was not administered until the beginning of week 5, we might expect two "phases," before and after vitamin E, making things more complicated.

Again, some formal framework for representing this situation and addressing the primary research question is required.

EXAMPLE 3: Growth of two different soybean genotypes. This study was conducted by Colleen Hudak, a former student in the Department of Crop Science at North Carolina State University, and is reported in Davidian and Giltinan (1995, p. 7). The goal was to compare the growth patterns of two soybean genotypes, a commercial variety, Forrest (F) and an experimental strain, Plant Introduction #416937 (P).

Data were collected in each of three consecutive years, 1988–1990. In each year, 8 plots were planted with F, 8 with P. Over the course of the growing season, each plot was sampled at approximate weekly intervals. At each sampling time, 6 plants were randomly selected from each plot, leaves from these plants were mixed together and weighed, and an ***average leaf weight per plant*** (g) was calculated.

In Figure 1.4, the data from the 8 F plots and 8 P plots for 1989 are depicted.

The primary objective of the study was

- To compare ***growth characteristics*** of the two genotypes.

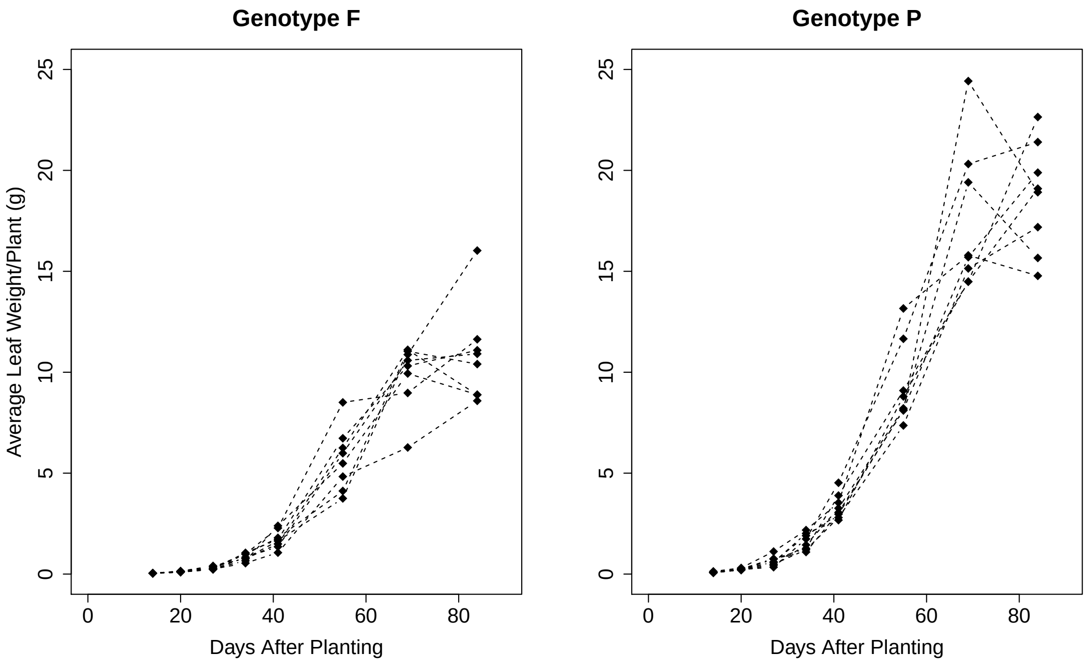

Figure 1.4: *Average leaf weight/plant profiles for 8 plots planted with Forrest and 8 plots planted with PI #416937 in 1989.*

From the figure, several features are notable:

- If we focus on the trajectory of a particular plot, we see that, typically, the growth begins ***slowly*** , with not much change over the first 3–4 observation times. Then, growth begins ***increasing*** at a faster rate in the middle of the season.
- Toward the end of the season, growth appears to begin ***leveling off***. This makes sense – soybean plants can only grow so large, so their leaf weight cannot increase without bound forever.
- Overall, then, the trajectory for any one plot does not appear to follow roughly a ***straight line*** as in the previous two examples, with an apparent constant rate of change over the observation period. Rather, the form of the trajectory seems more complicated, with almost an ***S shape***. It is thus clear that trying to characterize differences in growth characteristics will involve more than simply comparing a (constant) rate of change over the season.

In fact, the investigators expected that the growth pattern would not be as simple as an apparent straight line. They knew that growth would tend to level off toward the end of the season.

Thus, a more precise statement of the primary objective is

- To compare the ***limiting***, or ***asymptotic***, average leaf weight/plant between the 2 genotypes.
- To compare the way in which growth ***changes*** during the middle of the growing season.
- To compare the apparent ***initial*** average leaf weight/plant.

Several ***theoretical models*** have been postulated to describe growth processes exhibiting features like those observed for the soybean data. The most common such model is the ***logistic growth model***, which says that ***growth rate relative to present size*** decreases ***linearly*** with increasing size. Formally, letting Y denote the growth value (average leaf weight/plant here) and t denote time, this may be expressed by the ***deterministic relationship***

$$\frac{dY}{dt}\bigg/ Y = k\left(1 - \frac{Y}{a}\right),$$

where the right hand side is a linear function of present size Y, and $k > 0$ and $a > 0$. Upon integration, this model leads to

$$Y = \frac{a}{1 + c\exp(-kt)}, \tag{1.1}$$

where c is the value such that $a/(1+c)$ represents growth value at time $t = 0$. As $t \to \infty$, $Y \to a$, so that a is a ***physically meaningful*** parameter characterizing ***asymptotic behavior***, and, together with c, it characterizes the physically meaningful feature of “starting growth” at $t = 0$. The parameter $k > 0$ describes the ***change of growth with time***. A plot of the function (1.1) over a range of t for various choices of a, k, and c reveals that it has an ***S shape***.

This model appears to be a reasonable way to represent the growth profile for a ***given plot***. Figure 1.4 suggests that, although individual plot profiles have a similar shape, the parameters a and k describing asymptotic and change of growth might be ***different*** for different plots.

From Figure 1.4, it seems that average leaf weight/plant achieves “higher” limiting growth for genotype P relative to genotype F. That is, the ***asymptotic behavior*** seems to begin at lower values of the response for genotype F. The two genotypes seem to start off at roughly same value. It is difficult to make a simple statement about the relative rates of growth from the figure.

As it happened, ***weather patterns*** differed considerably over the three years: in 1988, conditions were unusually dry; in 1989, they were unusually wet; and in 1990 were relatively normal. Thus, comparison of growth patterns across the different weather patterns as well as how the weather patterns affect the comparison between genotypes, was also of interest.

Naturally, the investigators would like to be more formal about these observations and questions. This could be accomplished by incorporating a model like (1.1) within an appropriate statistical framework.

EXAMPLE 4: Pharmacokinetics of theopylline. A common objective is to investigate the ***pharmacokinetics*** (PK) of a drug; PK is, roughly speaking, the study of ***what the body does to the drug***. In a typical experimental PK study, a known dose of the drug is given to each of several (human or animal) subjects, and at several subsequent time points, blood samples are drawn from each and the ***concentration*** of drug in blood or plasma (the response) is determined for each sample.

The goal of such a study is

- To characterize the ***processes*** of drug ***absorption*** into the body, ***distribution*** throughout the body, and ***elimination*** from the body; the ***typical*** behavior of these processes; and how these processes ***vary*** in the population of subjects.

Armed with this information, ***pharmacokineticists*** develop ***dosing recommendations*** for the likely patient population taking the drug that will appear, e.g., in the ***labeling***.

Figure 1.5 shows concentration-time profiles for 4 of the 12 subjects in a PK study of the anti-asthmatic agent theophylline, each of whom received a single oral dose of theophylline at time 0 (given in units of mg/kg, so scaled to each individual's body weight in kg).

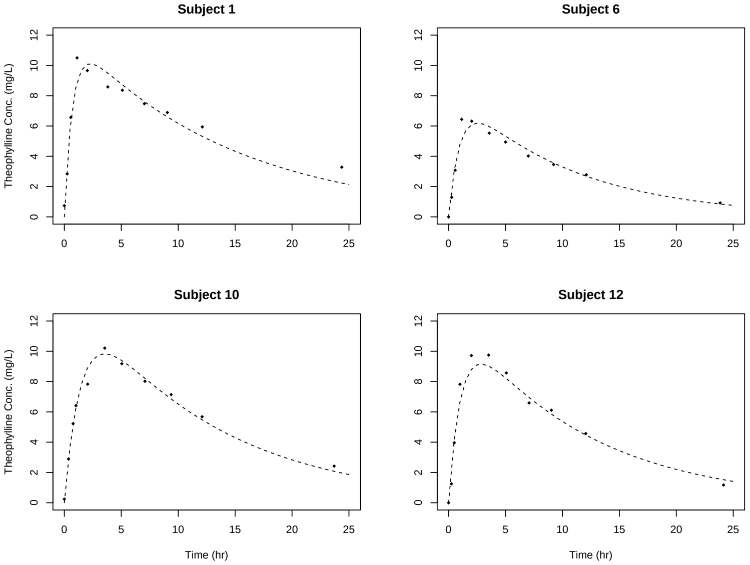

Figure 1.5: *Theophylline concentration-time profiles for 4 subjects receiving an oral dose of theophylline at time 0, with fits of model (*1.3) superimposed.

For each subject, 10 blood samples were drawn following the dose and assayed for theophylline concentration.

The profiles all exhibit the same general shape, with an early, steep rise in concentration that reaches a peak followed by what appears to be an ***exponential*** sort of decay. However, initial steepness, value and timing of the peak, and nature of the decay are ***different*** across subjects.

A statistician might be tempted to describe the concentration-time profile for an individual subject using a ***polynomial***. However, this is not a good approximation, nor is it a ***meaningful representation*** that addresses the questions of interest.

Instead, akin to the use of growth models like (1.1) in the soybean study, pharmacokineticists appeal to a deterministic ***theoretical representation*** based on representing the body as a system of ***compartments*** corresponding to components like "blood" and "deeper tissues." For orally-administered theophylline, a standard model is the ***one compartment open model with first order absorption and elimination***, represented pictorially as in Figure 1.6.

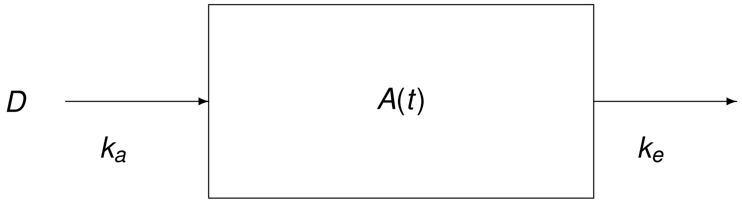

Figure 1.6: *One-compartment open model.*

In Figure 1.6, $A(t)$ is the amount of drug in the ***blood compartment*** at time t. Drug is ***absorbed*** through the gut into the blood at ***fractional absorption rate*** k_a and is ***eliminated*** (e.g., ***excreted*** by the kidneys and ***metabolized*** by the liver) at ***fractional elimination rate*** k_e.

The following ***system of differential equations***, with appropriate initial conditions, corresponds to Figure 1.6; see, for example, Gibaldi and Perrier (1982):

$$\begin{aligned} \frac{dA(t)}{dt} &= k_aA_a(t) - k_eA(t), \quad A(0) = 0 \\ \frac{dA_a(t)}{dt} &= -k_aA_a(t), \qquad\qquad A_a(0) = FD \end{aligned} \tag{1.2}$$

F = bioavailability, $A_a(t)$ = amount at absorption site

The system (1.2) can be solved for $A(t)$, and, dividing by V, the ***volume*** of the blood compartment, yields the following expression for ***concentration*** of drug at time t:

$$C(t) = \frac{A(t)}{V} = \frac{k_a DF}{V(k_a - k_e)}\{\exp(-k_e t) - \exp(-k_a t)\}, \; k_e = Cl/V, \tag{1.3}$$

The parameter F, ***bioavailability*** , is usually taken to be equal to 1. The parameter Cl, the ***clearance*** rate, is a measure of the volume of blood cleared of drug per unit time, and is of primary importance in understanding how the drug is eliminated from the system. The ***volume of distribution*** parameter V reflects the extent to which the drug distributes throughout the system.

The compartment model and the ensuing model for concentration $C(t)$ in (1.3) pertain to ***individual*** subject behavior; that is, the model is a theoretical description of biological processes taking place over time ***within*** a given subject, as that subject processes the drug. From Figure 1.5, which shows certain ***fits*** of the model (1.3) to each subject's data superimposed, it seems clear that differences in steepness, peak, and decay we noted earlier reflect the fact that each subject has his/her own PK parameters k_a, V, and Cl governing his/her individual PK behavior.

Returning to the objective of the study, characterizing ***typical*** absorption, distribution, and elimination behavior and how these processes ***vary*** in the population clearly will involve deducing the average values of the PK parameters and how they vary across the population from the concentration-time data. More formally, the goal is to describe the ***distribution*** of these parameters across the population.

An appropriate statistical framework is required in which the theoretical compartment model can be incorporated and within which this distribution can be characterized.

SO FAR: In the four examples we have considered, the outcome of interest is ***continuous*** in nature. That is,

- Distance (mm) from the center of the pituitary to the pterygomaxillary fissure
- Body weight (g)
- Average leaf weight/plant (g)
- Drug concentration (mg/L)

all can in principle take on any possible value in a particular range. How precisely we observe the value of the response is limited only by the precision of the measuring device we use.

In some situations, the outcome of interest is ***not*** continuous; rather, it is ***discrete*** in nature. We consider two additional examples.

EXAMPLE 5: Epileptic seizures and chemotherapy. A common situation is where the measurements are in the form of ***counts***. A response in the form of a ***count*** is by nature ***discrete*** – counts (usually) take only nonnegative integer values $(0, 1, 2, 3, ...)$.

The following famous data are reported by Thall and Vail (1990). A clinical trial was conducted in which 59 subjects with epilepsy suffering from simple or partial seizures were assigned at random to receive either the anti-epileptic drug progabide (subjects 29–59) or a ***placebo*** (an inert substance, subjects 1–28) in addition to a standard chemotherapy regimen all were taking. Because each individual might be prone to different rates of experiencing seizures, the investigators first tried to get a sense of this by recording the number of seizures suffered by each subject over the 8-week period prior to the start of administration of the assigned treatment. It is common in such studies to record such ***baseline*** measurements, so that the effect of treatment for each subject can be measured relative to how that subject behaved ***prior to*** treatment.

Following initiation of treatment, the number of seizures for each subject was counted for each of 4 consecutive 2-week periods. The age of each subject at the start of the study was also recorded, as it was suspected that subject age might be associated with the effect of the treatment.

The data for the first 5 subjects in each treatment group are summarized in Table 1.1.

	Period						
Subject	*1*	*2*	*3*	*4*	*Trt*	*Baseline*	*Age*
1	5	3	3	3	0	11	31
2	3	5	3	3	0	11	30
3	2	4	0	5	0	6	25
4	4	4	1	4	0	8	36
5	7	18	9	21	0	66	22
				⋮			
29	11	14	9	8	1	76	18
30	8	7	9	4	1	38	32
31	0	4	3	0	1	19	20
32	3	6	1	3	1	10	30
33	2	6	7	4	1	19	18

Table 1.1: *Seizure counts for 5 subjects assigned to placebo (coded as 0) and 5 subjects assigned to progabide (coded as 1).*

The primary objective of the study was to

- Determine if progabide ***reduces*** the rate of seizures relative to placebo in subjects like those in the trial.

We have repeated measurements (counts) on each subject over four consecutive observation periods, and we would like to compare somehow the baseline seizure counts to post-treatment counts, where the latter are observed ***repeatedly*** over time following initiation of treatment. Clearly, an appropriate analysis would make the best use of this feature of the data in addressing the main objective.

Note that some of the counts are quite small; in fact, for some subjects, 0 seizures (none) were experienced in some periods. For example, subject 31 in the treatment group experienced only 0, 3, or 4 seizures over the 4 observation periods. Clearly, models and methods that are appropriate for ***continuous*** outcomes like those in the first four examples would be suspect in this situation.

A classical approach to handling data in the form of counts is to ***transform*** them to some other scale. The motivation is to make them seem more ***normally distributed*** with constant variance, and the ***square root*** transformation is used to (hopefully) accomplish this. The desired result is that methods that are usually used to analyze continuous measurements can then be applied.

The drawback of this approach is that one is no longer working with the data on the ***original scale*** of measurement, numbers of seizures in this case. The statistical models assumed by this approach describe “square root number of seizures,” which is not particularly intuitive. Statistical methods that are designed to address questions on the basis of ***discrete*** repeated measurements like counts are required.

EXAMPLE 6: Maternal smoking and child respiratory health. Another common ***discrete data*** situation is where the outcome is ***binary***; that is, the outcome can take on only ***two*** possible values, which usually correspond to

- ***success*** or ***failure*** of a treatment to elicit a desired response
- ***presence*** or ***absence*** of some condition

The following data come from a very large public health study called the ***Six Cities Study***, which was undertaken in six small American cities to investigate a variety of public health issues. The full situation is reported in Lipsitz, Laird, and Harrington (1992). The portion we consider was focused on the association between maternal smoking and child respiratory health. Each of 300 children was examined once a year at ages 9–12. The outcome of interest was "***wheezing status***," a measure of the child's respiratory health, which was coded as either "no" (0) or "yes" (1), where "yes" corresponds to respiratory problems. Also recorded at each examination was a code indicating the mother's current level of smoking: 0 = none, 1 = moderate, 2 = heavy.

The data for the first 5 subjects are summarized in Table 1.2.

		Smoking at age				*Wheezing at age*			
Subject	*City*	9	10	11	12	9	10	11	12
1	Portage	2	2	1	1	1	0	0	0
2	Kingston	0	0	0	0	0	0	0	0
3	Portage	1	0	0	.	0	0	0	.
4	Portage	.	1	1	1	.	1	0	0
5	Kingston	1	.	1	2	0	.	0	1

Table 1.2: *Data for 5 children in the Six Cities study. Missing data are denoted by a "."*

The objective of an analysis of these data was to

- Determine how the ***typical wheezing response pattern*** changes with age
- Determine if there is an ***association*** between maternal smoking severity and child respiratory status (as measured by wheezing).

This study exemplifies the very common situation in medical and public health (and other) research of ***repeated measurements*** on a ***binary*** outcome. As with the count data, one might first think about trying to summarize and transform the data to allow (somehow) methods for continuous data to be used; however, this would clearly be inappropriate. As with the seizure study, statistical methods that are designed to address questions on the basis of ***discrete*** repeated measurements are required to address the questions of interest.

With binary outcome, spaghetti-type plots of repeated measures on the subjects as a function of age do not lend much insight. Informal inspection of individual subject data does suggest a possible association between wheezing and maternal smoking; e.g., subject 5 did not exhibit positive wheezing status until his/her mother's smoking increased in severity.

This highlights the fact that this situation is complex: over time (measured here by age of the child), an important characteristic, maternal smoking, ***changes***. Contrast this with the previous examples, where a main focus is to compare groups whose membership stays ***constant*** over time.

Another feature of these data is the fact that some data are ***missing*** for some subjects. Specifically, although the intention was to collect data for each of the four ages, this information is not available for some children and their mothers at some ages; for example, subject 3 has both the mother's smoking status and wheezing indicator missing at age 12. This pattern would suggest that the mother may have failed to appear with the child for this intended examination.

In the other examples, units (children, guinea pigs, plots, patients) were ***assigned*** to treatments; thus, these can be regarded as ***controlled experiments***, where the investigator has some control over how the factors of interest are "applied" or administered to the units (through ***randomization***). In contrast, in this study, the investigators did not decide which children would have mothers who smoke; instead, they could only ***observe*** smoking behavior of the mothers and wheezing status of their children. That is, this is an ***observational study***. Because it might be impossible or unethical to randomize subjects to potentially hazardous circumstances, studies of issues in public health and the social sciences are often ***observational***.

As in many observational studies, an additional difficulty is the fact that the ***exposure*** of interest, in this case maternal smoking, ***also changes*** with the response over time. This leads to complicated issues of interpretation in statistical modeling that are a matter of some debate. We discuss these issues in our subsequent development.

SUMMARY: These five examples illustrate the broad range of applications where data in the form of repeated measurements arise and the diverse range of questions of interest that are posed. The response of interest can be ***continuous*** or ***discrete***. The scientific questions might focus on very specific features of response trajectories, e.g., asymptotic behavior or PK processes; or might involve vague questions about the form of the ***typical response trajectory***. To complicate matters further, in studies where data were planned to be collected at certain points in time, it is possible for some responses to be ***missing***.

1.3 Statistical models for longitudinal data

In this course, we discuss a number of approaches for modeling data like those in the examples and describe different statistical methods for addressing questions of scientific interest within the context of these models.

STATISTICAL MODELS: Recall that a statistical model is a representation of the way in which data are thought to arise. Formally, a ***statistical model*** is a class of ***probability distributions*** that is assumed to have generated the data, where the data are represented by appropriately defined ***random variables***. Thus, a statistical model is a class of joint probability distributions for these random variables, which the analyst believes is a plausible representation of the true data generating mechanism.

The nature and features of an assumed statistical model dictate how questions of interest can be stated formally and unambiguously and how the data should be analyzed to address the questions. Different models embody different assumptions about how the data arise. Thus, the extent to which valid inferences on the questions of interest can be drawn under an assumed statistical model rests on how relevant its assumptions are to the situation at hand.

Thus, to appreciate the basis for techniques for data analysis and to use them appropriately, one must refer to and understand the associated statistical models. This connection is especially critical in the context of longitudinal data.

BASIC REPRESENTATION OF LONGITUDINAL RESPONSES: As in all of our examples, we consider a ***scalar response*** that is recorded on the same individual over time or some other set of conditions. In the specification of statistical models, it is convenient to think of all responses on the same individual ***together***, so that complex relationships among them can be summarized. Accordingly, we represent the responses at each time as ***random variables*** and collect these for each individual into ***random vectors***, as follows.

In general, define the random variable

$$Y_{ij} = \text{the } j\text{th response recorded on individual } i,$$

where $i = 1, \dots, m$ indexes individuals, and $j = 1, \dots, n_i$ indexes repeated measurements on the ith individual. Here, n_i, the number of repeated measurements on individual i, can be different for different individuals.

Let

$$t_{ij} = \text{the time at which } Y_{ij} \text{ is recorded for the } i\text{th individual.}$$

This notation allows further for the possibility of different times for different individuals.

For the ith individual, collect Y_{ij}, $j = 1, \dots, n_i$, into the random vector

$$\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T.$$

For example, consider the dental study data in Figure 1.2. If we index all children regardless of gender by i, then $i = 1, \dots, m = 27$. Each child was measured at ages 8, 10, 12, and 14 years; thus, $n_i = 4$ for all children, and $t_{i1} = 8$, $t_{i2} = 10$, and so on for all children in the study. The dental distances for the ith child can be summarized by the (4×1) random vector

$$\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{i4})^T.$$

We use this notation throughout to represent observed responses on each individual. In subsequent chapters, we will develop further notation to represent additional ***covariate*** information, such as gender in this example, dose in the theopylline pharmacokinetics study in ***EXAMPLE 4***, or maternal smoking status in the Six Cities study in ***EXAMPLE 6***, that may be available on each individual.

UNEQUAL n_i ***AND MISSING DATA:*** We emphasize that the foregoing notation is meant to refer to outcomes that are ***actually observed and recorded***. That is, $\boldsymbol{Y}_i$ is the $(n_i \times 1)$ vector of responses recorded at times $t_{i1}, \dots, t_{in_i}$ on individual i and ***available to the data analyst***.

In many settings, such as the dental study or the Six Cities study, the intention of the investigators is to collect outcomes from each individual at the ***same*** times (or other conditions). E.g., in the dental study, children were to be evaluated at ages 8, 10, 12, and 14; and mother-child pairs were to be assessed in the Six Cities study when the child was ages 9, 10, 11, and 12.

- In the dental study, all children were seen at the intended times. Accordingly, for child i, the random vector $\boldsymbol{Y}_i$ of outcomes actually ***observed*** is the same as the random vector of ***intended*** outcomes.
- In contrast, in the Six Cities study, the intended wheezing outcome for some mother-child pairs is ***missing***; for example, subject $i = 5$ is missing the outcome at age 10. In this case, $n_i = 3$, $t_{i1} = 9$, $t_{i2} = 11$, and $t_{i3} = 12$, and $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})^T$, with realized value $(0, 0, 1)^T$.

- In the sequel, when we discuss the implications of ***missing data***, we take care to distinguish the vector of ***intended*** outcomes from the vector of ***observed*** responses.

CLASSICAL NOTATION: Although summarizing repeated outcomes as random vectors for each individual is fundamental to modern longitudinal data models and methods, it was ***not*** generally used in the original formulation of ***classical models and methods*** for the analysis of (***continuous***) repeated measurements, namely, the ***univariate repeated measures analysis of variance*** methods we will discuss in Chapter 3. As we demonstrate, specification of these methods is through "***analysis of variance***" notation, where each scalar response is identified by subscripts corresponding to individual, group factors, and time; and time is viewed as a ***categorical factor*** rather than as potentially ***continuous***. The scalar responses for each individual are ***not*** collected into individual-specific random vectors in the standard presentation of the classical models.

We present classical models and methods using the ***original notation*** and then reexpress them in the notation given above so that they can be contrasted to the more modern methods.

CONTINUOUS RESPONSE: When the response is ***continuous***, not surprisingly, both classical and more modern methods are based on assuming each scalar response Y_{ij} (given covariate information) is ***normally distributed*** so that each vector $\boldsymbol{Y}_i$ (given covariate information) is taken to follow a ***multivariate normal distribution***. Moreover, models for the $\boldsymbol{Y}_i$ as a function of covariate information are often taken to be ***linear*** in parameters that characterize features of interest.

The methods we discuss in Chapters 3–6 are based on these assumptions. However, as we will demonstrate, even when normality does not hold, as long as m is "***large***," the methods have good properties that can be approximated via ***asymptotic theory***.

DISCRETE RESPONSE: When the outcome is ***discrete***, for example, binary or in the form of a count, things become more complicated. For a scalar discrete random variable Y_{ij}, probability distributions such as the Poisson or Bernoulli are standard models. However, in contrast to the normal, these distributions ***do not*** have immediate ***multivariate generalizations***. In addition, usual regression models for such scalar responses as a function of covariates, such as ***loglinear*** or ***logistic*** models, are ***not*** linear in parameters.

As a result, taking an approach analogous to that for continuous repeated outcomes is not directly possible. This challenge led to an enormous body of work in the 1980s and 1990s on approaches to modeling repeated discrete responses and associated inferential methods. In Chapters 7–9, we discuss these modeling strategies and associated methods. These are also applicable in the case of continuous responses for which linear models are ***not appropriate*** , as in the soybean growth study in ***EXAMPLE 3*** and the pharmacokinetic study of ***EXAMPLE 4***.

1.4 Outline of the course

Given the considerations of the previous section, the course offers coverage of two main areas. First, methods for the analysis of continuous repeated measurements that are reasonably thought of as normally distributed and for which linear models may be appropriate are discussed. Later, methods for the analysis of repeated measurements that are not reasonably thought of as normally distributed, such as discrete outcomes, or for which linear models are not appropriate, are covered.

The course can be thought of as having five parts:

I. Preliminaries:

- Introduction and motivation (Chapter 1)
- Modeling longitudinal data (Chapter 2)

II. Classical methods:

- Repeated measures analysis of variance (Chapter 3)

III. Methods for continuous, normally distributed responses:

- Modern methods: preliminaries (Chapter 4)
- Population-averaged linear models for continuous responses (Chapter 5)
- Linear mixed effects models (Chapter 6)

IV. Methods for discrete responses and problems involving nonlinear models:

- Review of generalized linear and nonlinear models (Chapter 7)
- Population-averaged models and generalized estimating equations (Chapter 8)
- (Subject-specific) generalized linear and nonlinear mixed effects models (Chapter 9)

V. Advanced topics

It is important to stress that there are ***numerous approaches*** to the modeling and analysis of longitudinal data, and there is no strictly “***right*** ” or “***wrong*** ” way. It is true, however, that some approaches are more flexible than others, imposing less restrictions on the nature of the data and allowing questions of scientific interest to be addressed more directly. We discuss how various approaches compare as we proceed.

Missing data , often the result of ***dropout*** from studies in which individuals are to be evaluated at several time points, are a recurring challenge in longitudinal data analysis. Although this is not a course on analysis in the presence of missing data, because missing data are a fact of life in longitudinal studies, we discuss their implications.