# 1 Introduction and Motivation

## 1.1 Purpose of this course

*OBJECTIVE:* The goal of this course is to provide an overview of statistical models and methods that are useful in the analysis of **longitudinal data**; that is, data in the form of **repeated measurements** on the same **unit** (human, plant, plot, sample, etc.) over time.

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

- In agriculture, a measure of growth may be taken on the same plot weekly over the growing season. Plots are assigned to different treatments at the start of the season.
- In a medical study, a measure of viral load (roughly, amount of HIV virus present in the body) may be taken at monthly intervals on patients with HIV infection. Patients are assigned to take different treatments at the start of the study.

Note that a defining characteristic of these examples is that the **same** response is measured repeatedly on each unit; i.e. viral load is measured again and again on the same subject. This particular type of data structure will be 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 issues concerning 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 in order to address these questions. Complementing the models, specialized methods of analysis are required.

In this course, we will study ways to model these data, and we will explore both classical and more recent approaches to analyzing them. Interest in the best ways to represent and interpret longitudinal data has grown tremendously in recent years, and a number of new powerful statistical techniques have been developed. We will discuss these techniques in some detail.

*TERMINOLOGY:* Although the term **longitudinal** naturally 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 may be taken repeatedly on the same unit. For example,

- The units may be human subjects. For each subject, reduction in diastolic blood pressure is measured on several occasions, each occasion involving administration of a different dose of an anti-hypertensive medication. Thus, the subject is measured repeatedly over **dose**.
- The 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.
- The units may be pregnant female rats. Each rat 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 is a bit different from the other two in that there is no natural **order** to the repeated measurements.

Thus, the methods will apply more broadly than the strict definition of the term **longitudinal data** indicates – the term will mean, to us, data in the form of **repeated measurements** that may well be over time, but may also 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.

We will use the term **response** to denote the measurement of interest. Because units are often human or animal subjects, we use the terms **unit**, **individual**, and **subject** interchangeably.

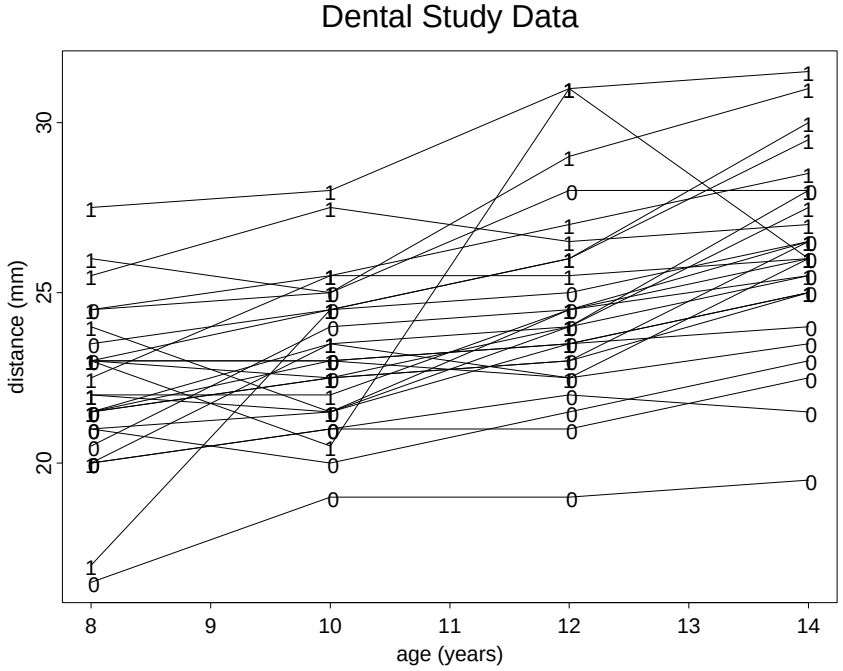
## 1.2 Examples

To put things into firmer perspective, we consider several real datasets from a variety of applications. These will not only provide us with concrete examples of longitudinal data situations, but will also serve to illustrate the range of ways that data may be collected and the types of measurements that may be of interest.

*EXAMPLE 1:* The orthodontic study data of Potthoff and Roy (1964).

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. In Figure 1, the distance measurements are plotted against age for each child. 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 may be seen.

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




Plots like Figure 1 are often called **spaghetti plots**, for obvious reasons!

The objectives of the study were to

- Determine whether distances over time are larger for boys than for girls
- Determine whether the rate of change of distance over time is similar for boys and girls.

Several features are notable from the plot of the data:

- It appears that 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.

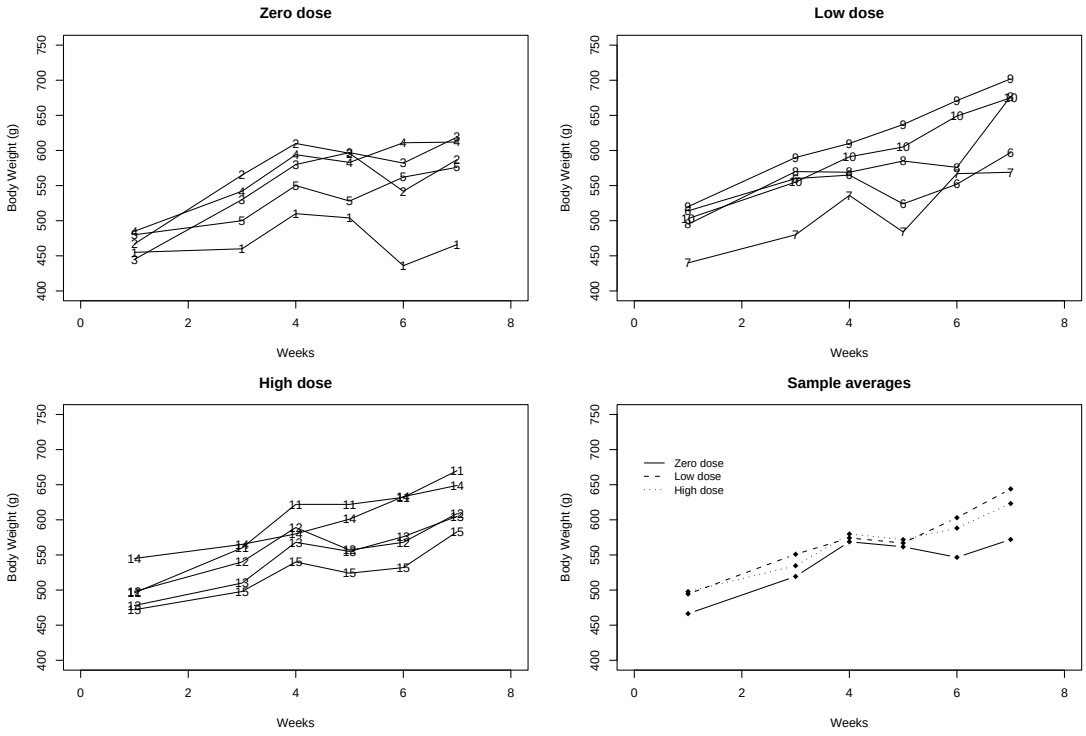
- Although boys seems to have larger distance measurements, the **rate of change** of the measurements with increasing age seems similar. More precisely, the **slope** of the increasing (approximate straight-line) relationship with age seems roughly similar for boys and girls. However, for any **individual** boy or girl, the rate of change (slope) may be steeper or shallower than the evident "typical" rate of change.

To address the questions of interest, it is clear that some formal way of representing the fact that each child has an individual-specific trajectory 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.

The following data are reported by Crowder and Hand (1990, p. 27) The study concerned the effect of a vitamin E diet supplement on the growth of guinea pigs. 15 guinea pigs were all 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 2, the data for the three dose groups are plotted on three separate graphs; the plotting symbol is the ID number (1–15) for each guinea pig. The plotting is similar to that for the dental data.

Figure 2: *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 whether the growth patterns differed 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.
- The trajectories for the zero dose group seem somewhat "lower" than those in the other dose groups.
- It is unclear whether 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 may 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.

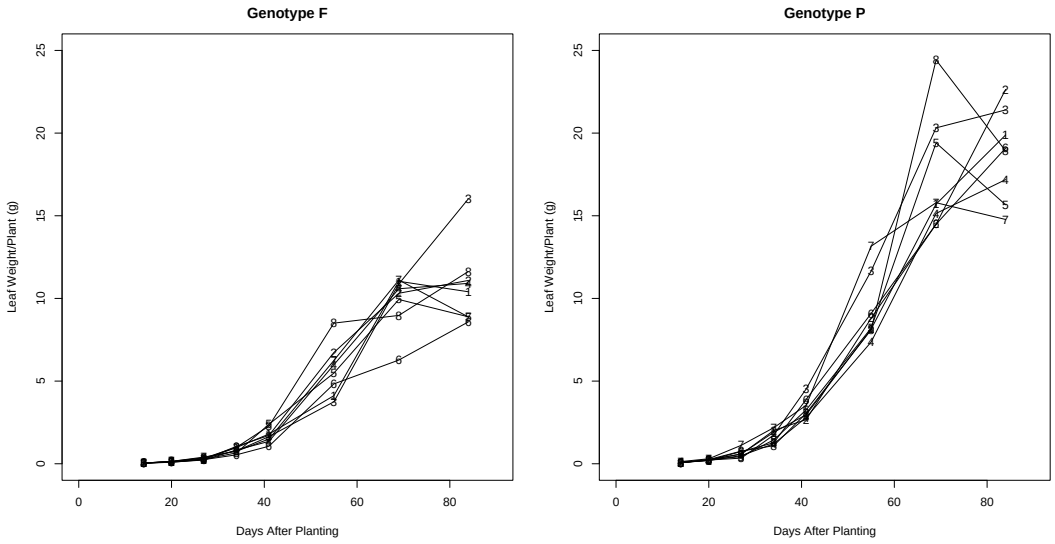
*EXAMPLE 3:* Growth of two different soybean genotypes.

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 weighted, and an average leaf weight per plant (g) was calculated. In Figure 3, the data from the 8 F plots and 8 P plots for 1989 are depicted.

The primary objective of the study was

- To compare the growth characteristics of the two genotypes.

Figure 3: *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 may 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 have the rough form of 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" type shape. It is thus clear that trying to characterize differences in growth characteristics will involve more than simply comparing rate of change over the season.

In fact, the investigators realized 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 their primary objective was

- To compare the apparent "limiting" average leaf weight/plant between the 2 genotypes.
- To compare the way in which growth accelerates during the middle of the growing season.
- To compare the apparent initial average leaf weight/plant.

From Figure 3, it seems that average leaf weight/plant achieves "higher" limiting growth for genotype P relative to genotype F. That is, the "leveling off" 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. Naturally, the investigators would like to be able to be more formal about these observations.

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

*SO FAR:* In the three examples we have considered, the measurement 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)

all may 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 response of interest is **not** continuous; rather, it is **discrete** in nature. That is, the values that we may observe differ by fixed amounts. For definiteness, we consider 2 additional examples:

*EXAMPLE 4:* 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 $(0, 1, 2, 3, \ldots)$.

The following data were first reported by Thall and Vail (1990). A clinical trial was conducted in which 59 people with epilepsy suffering from simple or partial seizures were assigned at random to receive either the anti-epileptic drug progabide (subjects 29–59) or an inert substance (a **placebo**, 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 may be measured relative to how that subject behaved before treatment.

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

The data for the first 5 subjects in each treatment group are summarized in Table 1.

Table 1: *Seizure counts for 5 subjects assigned to placebo (0) and 5 subjects assigned to progabide (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 |

The primary objective of the study was to

- Determine whether progabide reduces the rate of seizures in subjects like those in the trial.

Here, we have repeated measurements (counts) on each subject over four consecutive observation periods for each subject. Obviously, 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.

Moreover, 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, pretending that the response is **continuous** would be a lousy approximation to the true nature of the data! Thus, it seems that methods suitable for handling **continuous** data problems like the first three examples here would not be appropriate for data like these.

To get around this problem, a common 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 may then be applied.

However, 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 being assumed by this approach describe "square root number of seizures," which is not particularly interesting nor intuitive. Recently, new statistical methods have been developed to allow analysis of **discrete** repeated measurements like counts on the original scale of measurement.

*EXAMPLE 5:* Maternal smoking and child respiratory health.

Another common **discrete data** situation is where the response is **binary**; that is, the response may take on only **two** possible values, which usually correspond to things like

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

Clearly, it would be foolish to even try and pretend such data are approximately continuous!

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 current study 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 response 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 to indicate 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.

Table 2: *Data for 5 children in the Six Cities study. Missing data are denoted by a "."*

| | | *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 |

The objective of an analysis of these data was to

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

Note that it would be pretty pointless to plot the responses as a function of age as we did in the continuous data cases – here, the only responses are 0 or 1! Inspection of individual subject data does suggest that there is something going on here; for example, note that 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 situations, where a main focus is to compare groups whose membership stays constant over time.

Thus, we have **repeated measurements**, where, to further complicate matters, the measurements are **binary**! 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 we will see later in the course, methods for dealing with repeated binary responses and scientific questions like those above have been developed.

Another feature of these data is the fact that some measurements 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.

A final note: In the other examples, units (children, guinea pigs, plots, patients) were **assigned** to treatments; thus, these may be regarded as **controlled experiments**, where the investigator has some control over how the factors of interest are "applied" 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 example of an **observational study**. Because it may 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 thing 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 will 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 may arise. The response of interest may be **continuous** or **discrete**. The questions of interest may be focused on very specific features of the trajectories, e.g. "limiting growth," or may involve vague questions about the form of the "typical" trajectory.

## 1.3 Statistical models for longitudinal data

In this course, we will 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:* A statistical model is a formal representation of the way in which data are thought to arise, and the features of the model dictate how questions of interest may be stated unambiguously and how the data should be manipulated and interpreted to address the questions. Different models embody different assumptions about how the data arise; thus, the extent to which valid conclusions may be drawn from a particular model rests on how relevant its assumptions are to the situation at hand.

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

Formally, a statistical model uses *probability distributions* to describe the mechanism believed to generate the data. That is, responses are represented by a *random variables* whose probability distributions are used to describe the chances that a response takes on different values. How responses arise may involve many factors; thus, how one "builds" a statistical model and decides which probability distributions are relevant requires careful consideration of the features of the situation.

*RANDOM VECTORS:* In order to

- elucidate the assumptions made under different models and methods and make distinctions among them
- describe the models and methods easily

it is convenient to think of all responses collected on the same unit over time or other set of conditions **together**, so that complex relationships among them may be summarized.

Consider the random variable

$$Y_{ij} = \text{the } j\text{th measurement taken on unit } i.$$

To fix ideas, consider the dental study data in Figure 1. Each child was measured 4 times, at ages 8, 10, 12, and 14 years. Thus, we let $j = 1, \ldots, 4$; $j$ is indexing the number of times a child is measured. To summarize the information on **when** these times occur, we might further define

$$t_{ij} = \text{the time at which the } j \text{ measurement on unit } i \text{ was taken.}$$

Here, for all children, $t_{i1} = 8$, $t_{i2} = 10$, and so on for all children in the study. Thus, if we ignore gender of the children for the moment, the responses for the $i$th child, where $i$ ranges from 1 to 27, are $Y_{i1}, \ldots, Y_{i4}$, taken at times $t_{i1}, \ldots, t_{i4}$. In fact, we may summarize the measurements for the $i$th child even more succinctly: define the $(4 \times 1)$ **random vector**

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix}.$$

The components are random variables representing the responses that might be observed for child $i$ at each time point. Later, we will expand this notation to include ways of representing additional information, such as gender in this example.

The important message is that it is possible to represent the responses for the $i$th child in a very streamlined and convenient way for the purposes of talking about them all together. Each child $i$ has its own **vector** of responses $\boldsymbol{Y}_i$. It often makes sense to think of the data not just as **individual** responses $Y_{ij}$, some from one child, some from another according to the indices, but rather as **vectors** corresponding to children, **the units** – each unit has associated with it an entire vector of responses.

It is worth noting that this way of summarizing information is not always used; in particular, some of the classical methods for analyzing repeated measurements that we will discuss are usually not cast in these terms. However, as we will see, using this unified way of representing the data will allow us to appreciate differences among approaches.

This discussion demonstrates that it will be convenient to use **matrix notation** to summarize longitudinal data. This is indeed the case in the literature, particularly when discussing some of the newer methods. Thus, we will need to review elements of of matrix algebra that will be useful in describing the models and methods that we will use.

*PROBABILITY DISTRIBUTIONS:* Statistical models rely on **probability distributions** to describe the way in which the random variables invoved in the model take on their values. That is, probability distributions are used to describe the chances of seeing particular values of the response of interest.

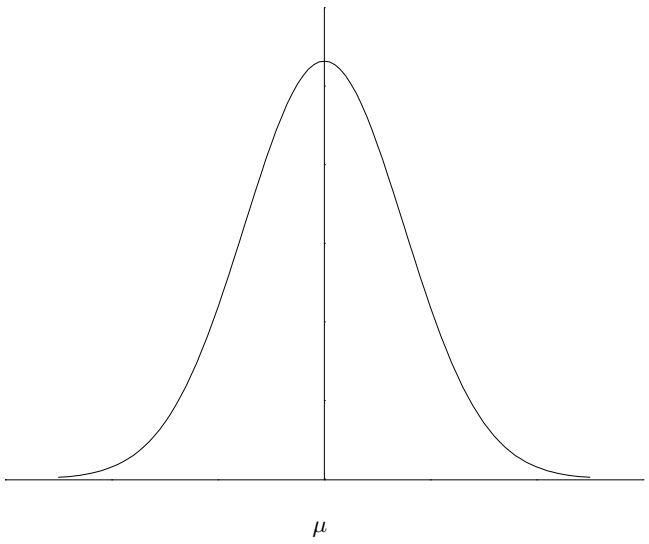
This same reasoning will of course be true for repeated measurements. In fact, acknowledging that it makes sense to think of the responses for each unit in terms of a **random vector**, it will be necessary to consider probability models for entire vectors of several responses thought of **together**, coming from the same unit.

*NORMAL DISTRIBUTION:* For **continuous** data, recall that the most common model for single observations is the **normal** or **Gaussian** distribution. That is, if $Y$ is a normal random variable with mean $\mu$ and variance $\sigma^2$, then the probabilities with which $Y$ takes on different values $y$ are described by the **probability density function**

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

This function is depicted graphically in Figure 4. Recall that the area under the curve between two values represents the probability of the random variable $Y$ taking on a value in that range.

Figure 4: *Normal density function with mean $\mu$.*




The assumption that data may be thought of as ending up the way they did according to the probabilities dictated by a normal distribution is a fundamental one in much of statistical methodology. For example, classical **analysis of variance** methods rely on the relevance of this assumption for conclusions (i.e. inferences based on $F$ ratios) to be valid. Classical methods for **linear regression modeling** also are usually motivated based on this assumption. When the response is continuous, the assumption of normality is often a reasonable one.

*MULTIVARIATE NORMAL DISTRIBUTION:* When we have data in the form of repeated measurements, we have already noted that it is convenient to think of the data from a particular unit $i$ as a **vector** of individual responses, one vector from each unit. We will be much more formal later; for now, consider that these vectors may be thought of as **unrelated** across individuals – how the measurements for one child turn out over time has nothing to do with how they turn out for another child. However, if we focus on a **particular** child, the measurements on that child will definitely be related to one another! For example, in Figure 1, the boy with the "highest" profile starts out "high" at age 8, and continues to be "high" over the entire period. Thus, we would like some way of not only characterizing the probabilities with which a child has a certain response at a certain age, but of characterizing how responses on the same child are related!

When the response is continuous and the assumption of normality seems reasonable, we will thus need to discuss the extension of the idea of the normal distribution from a model just for probabilities associated with a single random variable representing a response at one time to a model of the **joint** probabilities for several responses together in a random vector. This of course includes how the responses are related. The **multivariate normal distribution** is the extended probability model for this situation. Because many popular methods for the analysis of longitudinal data are based on the assumption of normally distributed responses, we will discuss the multivariate normal distribution and its properties in some detail.

*NORMAL, CONTINUOUS RESPONSE:* Armed with our understanding of matrix notation and algebra and the multivariate normal distribution, we will study methods for the analysis of continuous, longitudinal data in the first part of the course that are appropriate when the multivariate normal distribution is a reasonable probability model.

*DISCRETE RESPONSE:* Of course, the normal distribution is appropriate when the response of interest is **continuous**, so, although the assumption of normality may be suitable in this case, it may not be when the data are in the form of small counts, as in the seizure example. This assumption is certainly not reasonable for binary data. As discussed above, a common approach has been to try to transform data to make them "approximately normal" on the transformed scale; however, this has some disadvantages.

In the early 1980's, there began an explosion of research into ways to analyze **discrete** responses that did not require data transformation to induce approximate normality. These methods were based on more realistic probability models, the **Poisson** distribution as a model for **count** data and the **Bernoulli** (binomial) distribution as a model for **binary** data.

For regression-type problems, where a single response is measured on each unit, the usual classical linear regression methods were extended to allow the assumption that these distributions, rather than the normal distribution, are sensible probability models for the data. The term **generalized linear models** is used to refer to the models and techniques used.

Starting in the late 1980's, generalized linear model methods were **extended** to the situation of **repeated measurement** data, allowing one to think in terms of **random vectors** of responses, each element of which may be thought of as Poisson or Bernoulli distributed. We will study these probability distributions, generalized linear models, and their extension to longitudinal data.

*NONNORMAL, CONTINUOUS RESPONSE:* In fact, although the normal distribution is by far the most popular probability model for continuous data, it is not always a sensible choice. As can be seen from Figure 4, the normal probability density function is **symmetric**, saying that probabilities of seeing responses smaller or larger than the mean are the same. This may not always be reasonable.

As we will discuss later in the course, other probability models are available in this situation. It turns out that the methods in the same spirit as those used for discrete response may be used to model and analyze such data.

## 1.4 Outline of the course

Given the considerations of the previous section, the course will offer coverage of two main areas. First, methods for the analysis of continuous repeated measurements that are reasonably thought of as normally distributed will be discussed. Later, methods for the analysis of repeated measurements that are not reasonably thought of as normally distributed, such as discrete responses, are covered.

The course may be thought of as coming in roughly five parts:

**I. Preliminaries:**

- Introduction
- Review of matrix algebra
- Random vectors, multivariate distributions as models for repeated measurements, multivariate normal distribution, review of linear regression
- Introduction to modeling longitudinal data

**II. Classical methods:**

- Classical methods for analyzing normally distributed, balanced repeated measurements – "univariate" analysis of variance approaches
- Classical methods for analyzing normally distributed, balanced repeated measurements – "multivariate" analysis of variance approaches
- Discussion of classical methods – drawbacks and limitations

**III. Methods for unbalanced, normally distributed data:**

- General linear models for longitudinal data, models for correlation
- Random coefficient models for continuous, normally distributed repeated measurements
- Linear mixed models for continuous, normally distributed repeated measurements

**IV. Methods for unbalanced, nonnormally distributed data:**

- Probability models for discrete and nonnormal continuous response, generalized linear models
- Models for discrete and nonnormal continuous repeated measurements – generalized estimating equations

**V. Advanced topics:**

- Generalized linear mixed models for discrete and nonnormal continuous repeated measurements
- More general nonlinear mixed models for all kinds of repeated measurements
- Issues associated with missing data

Throughout, we will devote considerable time to the use of standard statistical software to implement the methods. In particular, we will focus on the use of the SAS (Statistical Analysis System) software. Some familiarity with SAS, such as how to read data from a file, how perform simple data manipulations, and basic use of simple procedures such as `PROC GLM` is assumed.

The examples in subsequent chapters are implemented using Version 8.2 of SAS on a SunOs operating system. Features of the output and required programming statements may be somewhat different when older versions of SAS are used, as some of the procedures have been modified. In addition, slight numerical differences arise when the same programs are run on other platforms. The user should consult the documentation for his/her version of SAS for possible differences.

Plots in the figures are made with R and Splus. Making similar plots with SAS is not demonstrated in these notes, as it is assumed the user will wish to use his/her own favorite plotting software.

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 will note how various approaches compare as we proceed.

Throughout, we adopt a standard convention. We often use upper case letters, e.g., $Y$ and $\boldsymbol{Y}$, to denote random variables and vectors, most often those corresponding to the response of interest. We use lower case letters, e.g., $y$ and $\boldsymbol{y}$, when we wish to refer to **actual data values**, i.e., **realizations** of the random variable or vector.