STATISTICAL MODELING AND INFERENCE

An Introduction and Overview

Marie Davidian

Department of Statistics

NC STATE UNIVERSITY

davidian@stat.ncsu.edu

http://www.stat.ncsu.edu/∼davidian

SAMSI Inverse Problem Workshop

September 21, 2002

OUTLINE

1. **Introduction:** *Sources of variation in data*
2. **Whirlwind review of probability**
3. **Statistical models**
4. **Statistical inference:** *Classical frequentist paradigm*
5. **Modeling and inference for independent data**
6. **Hierarchical statistical models for complex data structures**
7. **Statistical inference:** *Bayesian paradigm*
8. **Hierarchical models, revisited**
9. **Closing remarks**
10. **Where to learn more...**

1. Introduction

Statistics:

- "*The study of variation*"
- *NOT* a branch of mathematics
- More like a *philosophy* for thinking about *drawing conclusions from data*

Objective of this tutorial: Introduce *statistical thinking* from the point of view of *fitting models to data*

- Lay groundwork for further study

Deterministic models: Representation of an "*exact*" relationship

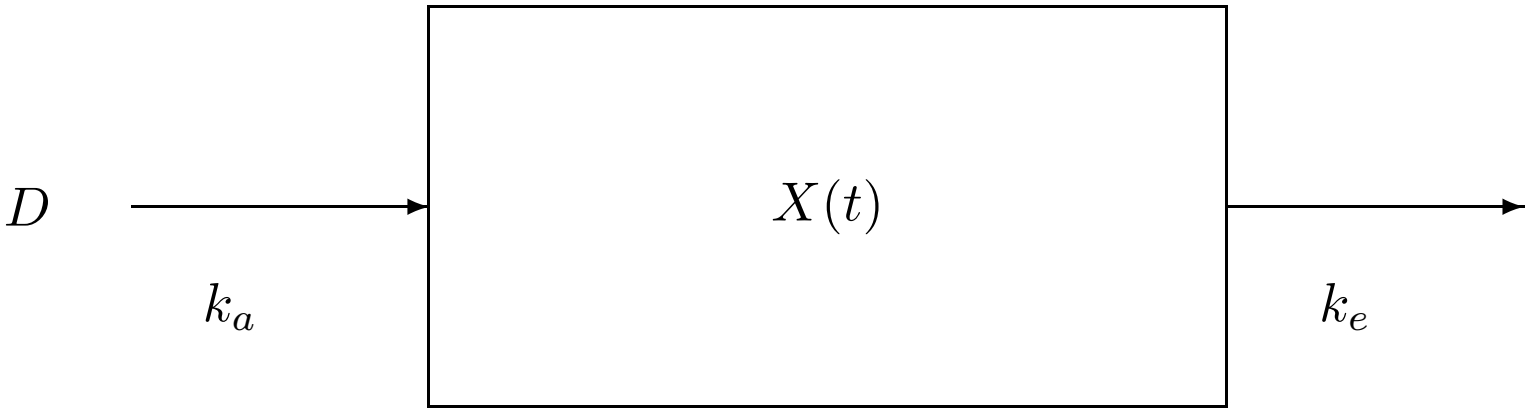
- *System* $\dot{x}(t) = g\{t, x(t), \theta\}$
- *Solution* $y = x(t, \theta)$
- *Objective:* "*Inverse problem*" – Learn about θ
- *Example* – the design problem (Banks)

Observations: Suppose we observe the system over time and record values $y_1, \ldots, y_n$ at times $0 \leq t_1 < \cdots < t_n$

- *Commonly*, observations *do not* track exactly on the curve $y = x(t, \theta)$

Simple example: *Pharmacokinetics of theophylline* (anti-asthmatic)

- Understanding of processes of *absorption*, *distribution*, *elimination* important for developing dosing recommendations
- Common deterministic model: *One compartment model with first-order absorption and elimination* following oral dose D



- *Assumption*: $X(t) = VC(t)$ [constant relationship between drug concentration in plasma $C(t)$ and amount of drug in body $X(t)$]

System: $X_a(t) =$ amount of drug at absorption site at time t

$$\begin{aligned} \dot{X}(t) &= k_a X_a(t) - k_e X(t) \\ \dot{X}_a(t) &= -k_a X_a(t) \end{aligned}$$

with initial conditions $X_a(0) = X_{a0} = FD$, $X(0) = X_0 = 0$, where F is the fraction available (assume known)

Closed form solution: Divide by V

$$C(t) = \frac{k_a F D}{V(k_a - k_e)}\{e^{-k_e t} - e^{-k_a t}\}$$

Result: If the model is a *perfect representation* of the system, the relationship

$$C(t) = \frac{k_a F D}{V(k_a - k_e)}\{e^{-k_e t} - e^{-k_a t}\}$$

should describe the concentration observed at time t

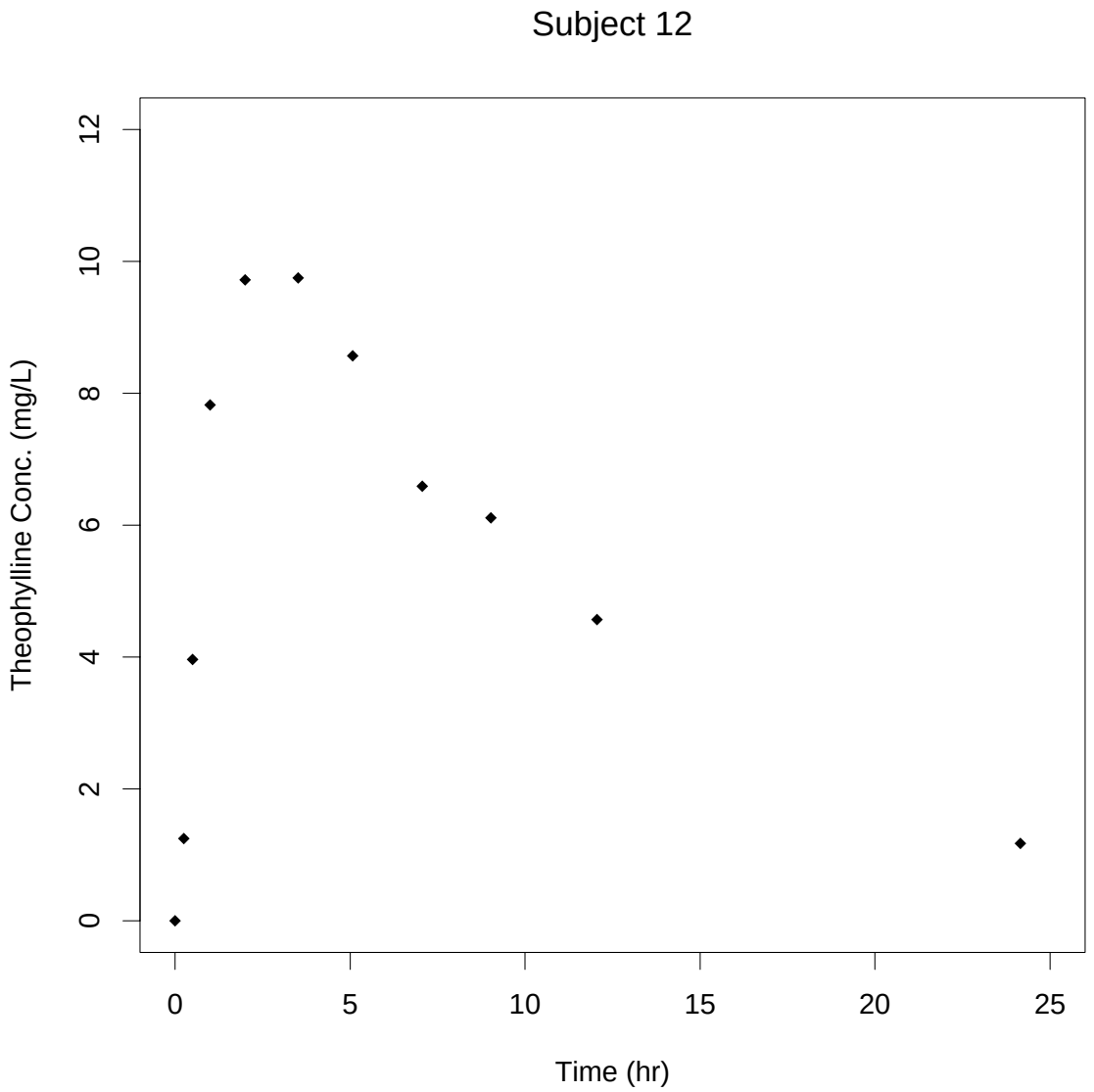
Experiment: PK in *humans* following oral dose

- 12 "*healthy volunteers*" each given dose D (mg/kg) at time $t = 0$
- Blood samples drawn at 10 subsequent time points over the next 25 hours for each subject
- Samples *assayed* for theophylline concentration
- *Observe* $y_1, \ldots, y_{10}$ at times $t_1, \ldots, t_{10}$

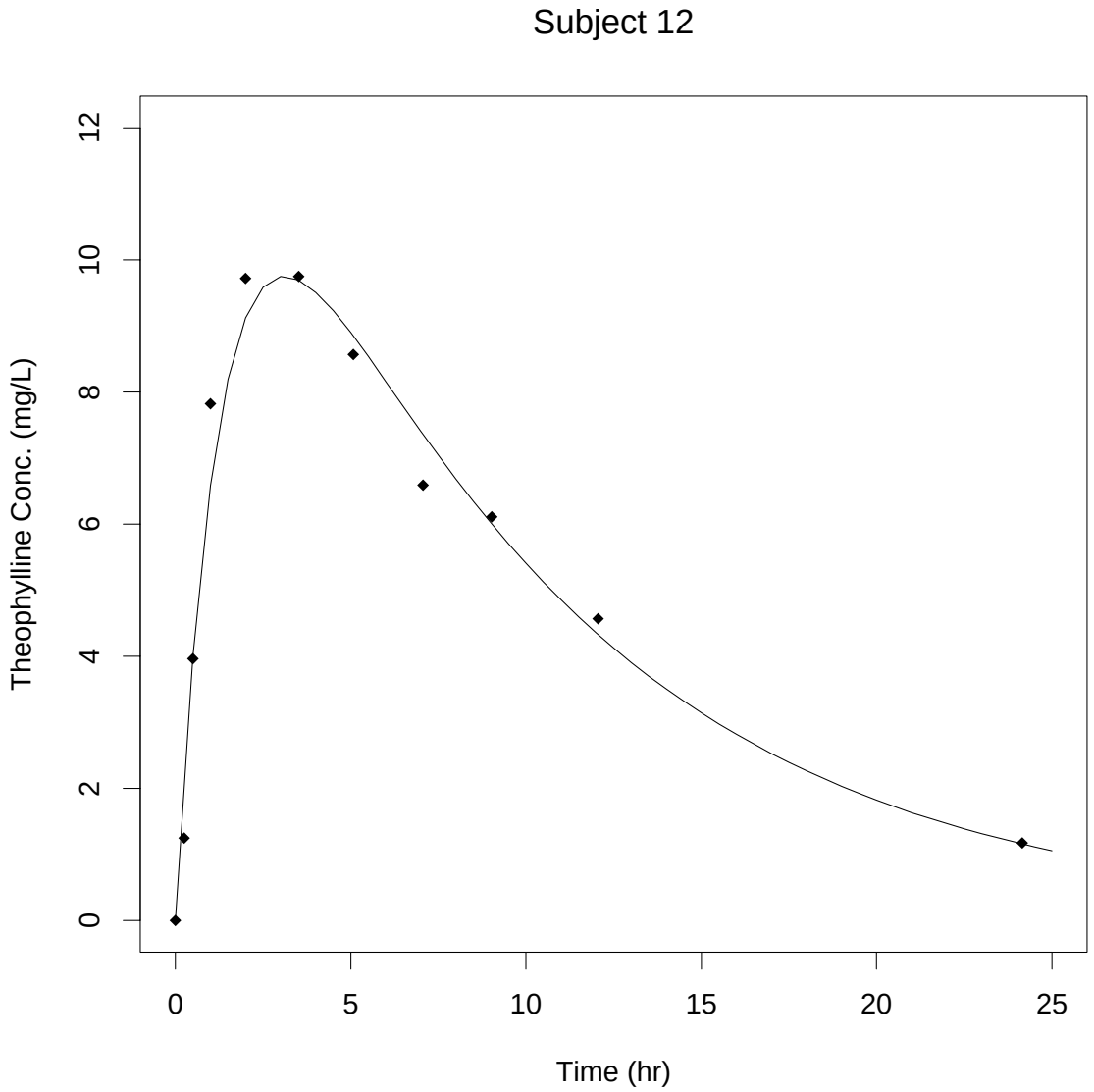
Objectives:

1. For a *specific subject*, learn about absorption, elimination, distribution by determining k_a, k_e, V $\Rightarrow$ dosing recommendations for this subject
2. Learn about how absorption, elimination, and distribution differ from subject to subject $\Rightarrow$ dosing recommendations for the *population* of likely subjects

Data for subject 12: Plot of concentration vs. time



Data for subject 12: With "fitted model" superimposed



Remarks:

- Observed concentrations trace out a pattern over time quite *similar* to that dictated by the one compartment model
- But they do not lie *exactly* on a smooth trajectory
- "*Observation error*"

Why?

- One obvious reason: Assay is not perfect, cannot measure concentration *exactly* (measurement error)
- Other reasons?

Remarks:

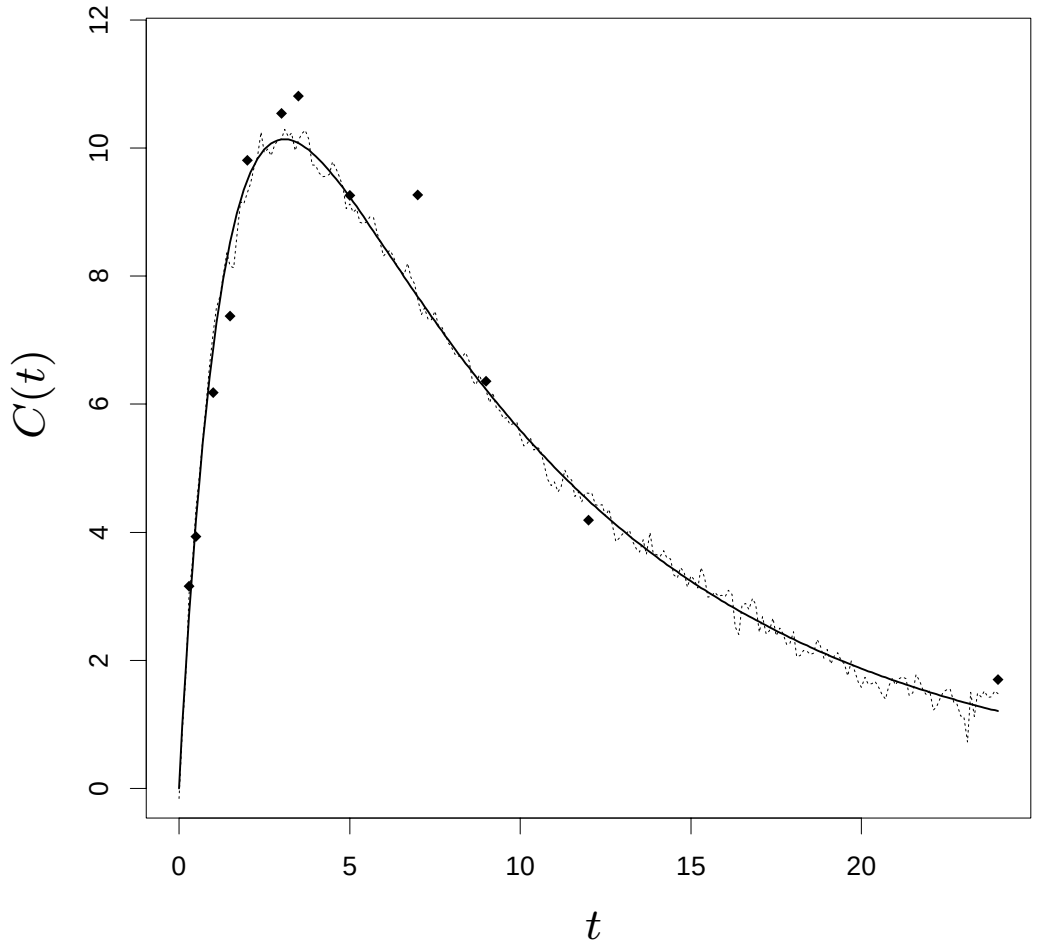
- Observed concentrations trace out a pattern over time quite similar to that dictated by the one compartment model
- But they do not lie *exactly* on a smooth trajectory
- "*Observation error*"

Why?

- One obvious reason: Assay is not perfect, cannot measure concentration *exactly* (measurement error)
- Other reasons?
  - Model *misspecification*
  - More *complex biological process*
  - Times/dose recorded *incorrectly*
  - Etc. . .

Hypothetically, what's really going on: More *complex biological process* and *measurement error*

- The model is an *idealized* representation in some sense



Sources of variation: The (*deterministic*) *model* is a good representation of the *general pattern*, but *observed concentrations* are subject to

- Intra-subject "*fluctuations*"
- Assay *measurement error*

Conceptualization: Can think of what we *observe* as

$$y_j = f(t_j, \theta) + \epsilon_j$$

- $f(t, \theta) = C(t)$, a function of $\theta = (k_a, k_e, V)^T$ (and D)
- ϵ_j is the *deviation* between what the (deterministic) model dictates we would see at t_j and what we actually observe due to *measurement error*, "*biological fluctuations*"

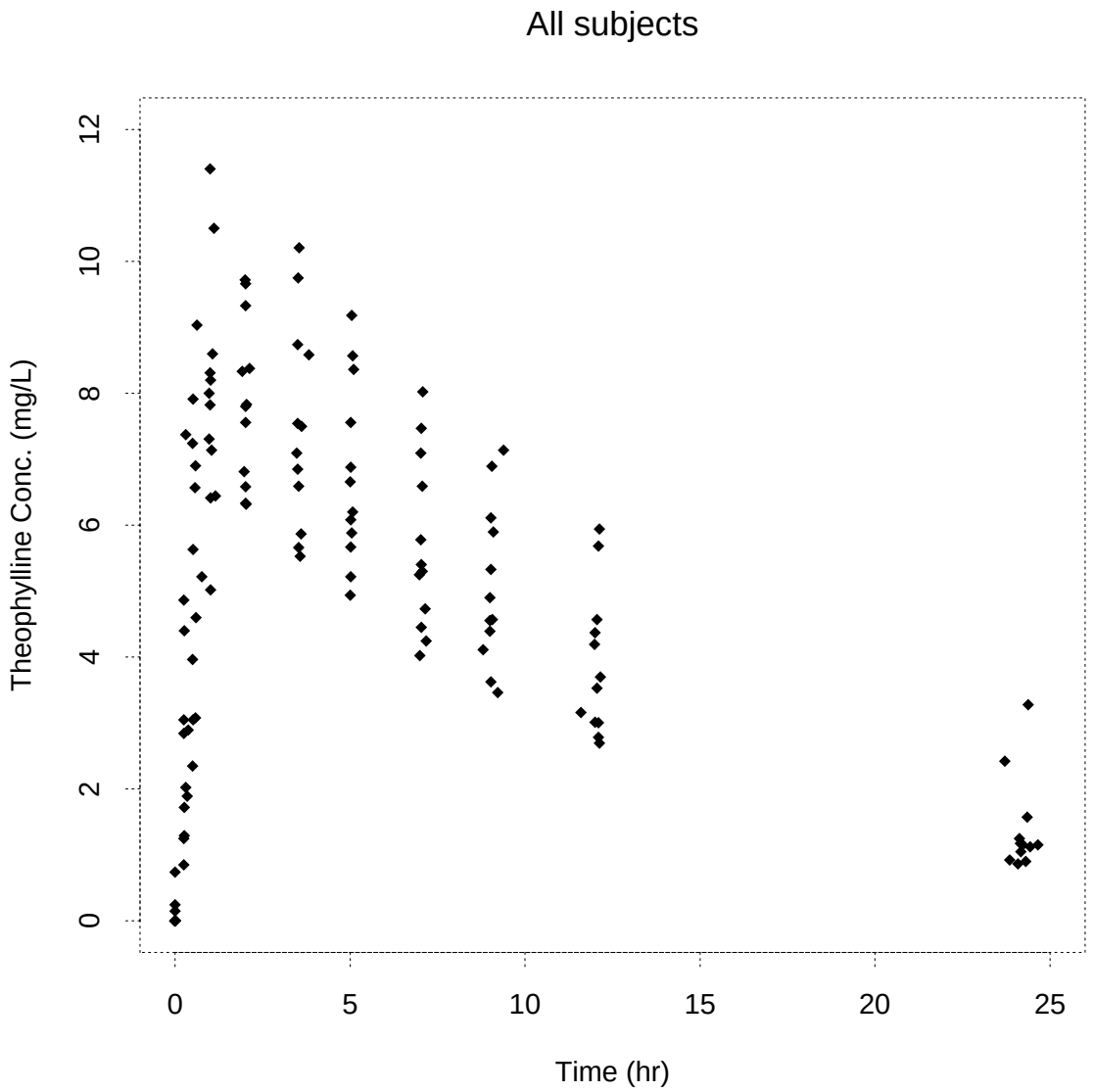
Thought experiment: Consider *measurement error*

- A particular blood sample has a "*true*" concentration
- When we measure this concentration, an error is committed, which causes "*observed*" to deviate from "*true*" ($+$ or $-$)
- Suppose we were to measure the same sample *over and over* – each time, a possibly different error is committed
- Thus, all such observations would turn out *differently* (*VARY*), even though, *ideally*, they should be all the *same* (measuring the *same thing*)

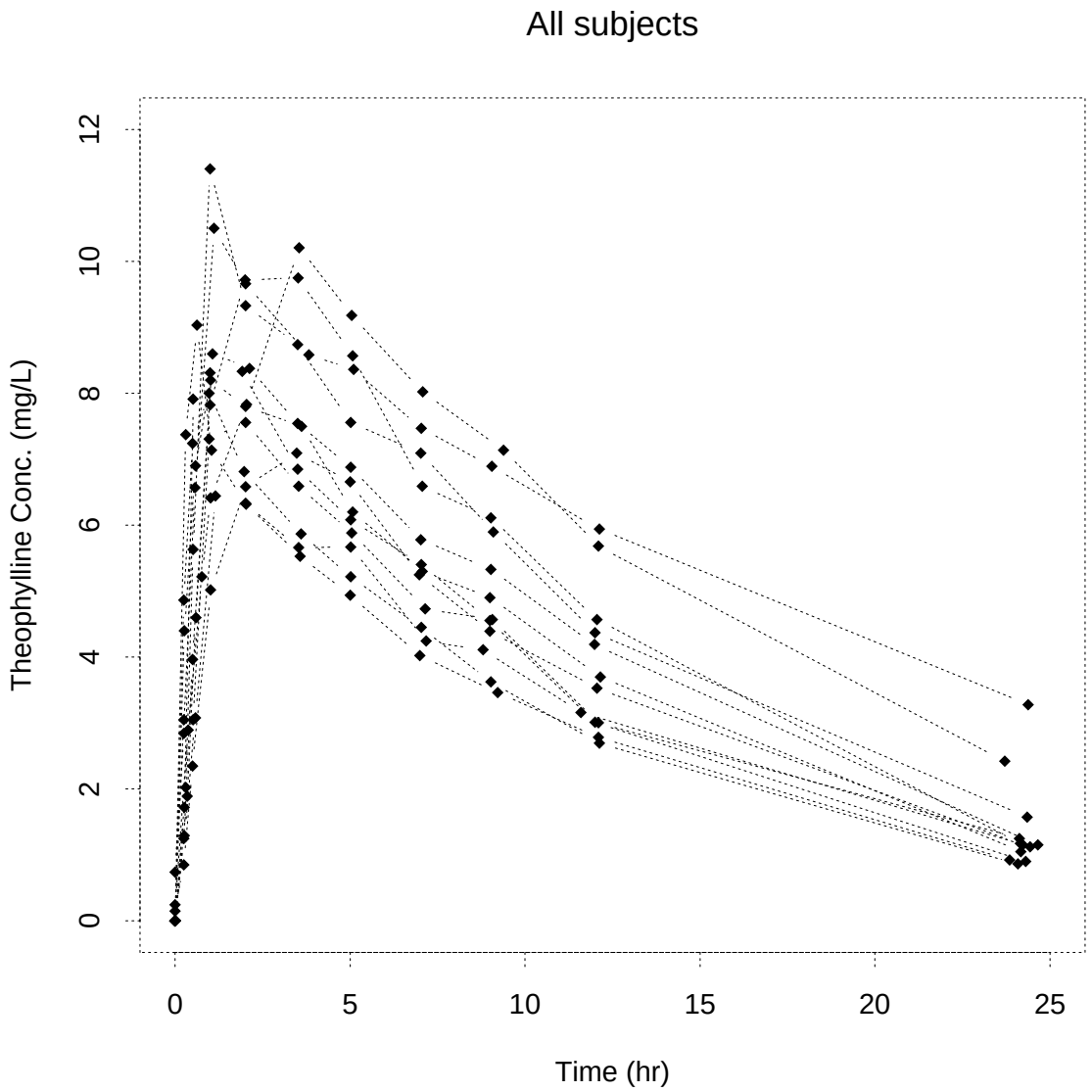
Result: Measurement error is a *source of variation* that leads to *UNCERTAINTY* in what we *observe*

- In actuality, we measure the concentration only *once*, but it could have *turned out differently*
- $\Rightarrow$ Any determination of θ from data is subject to *uncertainty*

All 12 subjects:



All 12 subjects:



Recall objective 2: Absorption, distribution, elimination in the *population* of subjects

- *Similar pattern* for all subjects, but *different features*
- $\Rightarrow$ Each subject has his/her *own* θ
- *Subject-to-subject variation*
- *Objective, restated* – learn about θ values in the (hypothetical) *population* of subjects like these
- But we have only seen a *sample* of 12 subjects from this population $\Rightarrow$ *uncertainty* about *entire* population
- *Each* subject's data *also* subject to *uncertainty* due to *measurement error*, "*biological fluctuation*"
- How to *formalize* the objective and take into account *uncertainty* from all these *sources of variation*?

Principles:

- Failure to acknowledge *uncertainty* can lead to *erroneous conclusions*
- Acknowledging *uncertainty* requires a *formal framework* to describe and assess it
- Acknowledging *uncertainty* clarifies *limitations* of what can be learned from *data*

Statistical models:

- Formally represent *sources of variation* leading to *uncertainty*
- $\Rightarrow$ In *this* context: Incorporate a *deterministic model* in a *statistical framework*
- Main tool: *probability*

For example: *Statistical model* for theophylline concentrations for subject 12

$$Y_j = f(t_j, \theta) + \epsilon_j, \quad j = 1, \ldots, n$$

- ϵ_j (and hence Y_j) is a *random variable* with a *probability distribution* that characterizes "*populations*" of possible values of phenomena like measurement errors, fluctuations that might occur at t_j
- Describes pairs (Y_j, t_j) we *might see* – describes the "*data generating mechanism*"
- *Data* we *observe* are realizations of Y_j, $j = 1, \ldots, n$: $y_1, \ldots, y_n$
- The *mechanism* is characterized by *assumptions* on the *probability distribution* of ϵ_j (so, equivalently, on that of Y_j)

2. Review of probability

"Experiment": *Sample space* Ω

- Toss a coin 2 times , $\Omega = \{HH, HT, TH, TT\}$
- Measure concentration, $\Omega = \{$ all possible conditions $\}$

Probability function: For $\mathcal{B}$ some collection of subsets A of Ω

- $P(A) \geq 0$, $P(\Omega) = 1$, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
- Properties – $P(\emptyset) = 0$, $P(A) \leq 1$, $P(A^c) = 1 - P(A)$, $P(A) \leq P(B)$ if $A \subset B$, etc

Random variable: A function from Ω into $\Re$ (capital letters) with *new sample space* $\mathcal{X}$

- E.g., Toss coin 2 times, $X(\omega) = \#$ heads, $\mathcal{X} = \{0, 1, 2\}$
- E.g., Measure concentration, $\epsilon(\omega)$ error committed, $\mathcal{X} = (-\infty, \infty)$

Probability function for X:

- $X = x \in \mathcal{X}$ iff $\omega \in \Omega \ni X(\omega) = x$

$$P_X(X = x) = P(\omega \in \Omega : \ X(\omega) = x)$$

- Customary to speak directly about probability wrt *random variables*
- X denotes *random variable*, x denotes possible values (elements of $\mathcal{X}$, *realizations*)

Cumulative distribution function for X: $F(x) = P(X \leq x)\ \forall x$ (*Nondecreasing* and *right continuous*)

- X is *discrete* if $F(x)$ is a step function
- X is *continuous* if $F(x)$ is continuous

Probability mass and density functions:

- Discrete: *probability mass function*

$$f(x) = P(X = x)\ \forall x \Rightarrow F(x) = \sum_{u \leq x} f(u)$$

- Continuous: *probability density function* $f(x)$ satisfies

$$F(x) = \int_{-\infty}^{x} f(u)\, du\ \forall x$$

- $f(x) \geq 0$, $\sum_x f(x) = 1$ or $\int_{-\infty}^{\infty} f(x)$

Transformations: $Y = g(X)$ is *also* a random variable with pmf/pdf that may be derived from $f(x)$

$\Rightarrow$ "*Probability distribution*"

Random vectors: $(X_1, \ldots, X_p)^T$ is a function from Ω into $\Re^p$ with pmf/pdf $f(x_1, \ldots, x_p)$

- E.g., all *discrete* – $f(x_1, \ldots, x_p) = P(X_1 = x_1, \ldots, X_p = x_p)$
- *Marginal pmf/pdf* – e.g., $f_{X_1}(x_1) = \sum_{x_2,\ldots,x_p} f(x_1, \ldots, x_p)$

Independence: X_1 and X_2 are *independent* if

$$f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2), \quad X_1 \perp\!\!\!\perp X_2$$

Expectation (mean, expected value): "*Average value*"

$$E\{g(X)\} = \begin{cases} \int_{-\infty}^{\infty} g(x) f(x)\, dx & X \text{ continuous} \\ \sum_x g(x) f(x) = \sum_x g(x) P(X = x) & X \text{ discrete} \end{cases}$$

Variance: Second central moment, measure of "*spread*," quantifies *variation*

$$\text{var}(X) = E\left[\{X - E(X)\}^2\right]$$

- Standard deviation $= \sqrt{\text{var}(X)}$ on same scale of X

Covariance and correlation: "*Degree of association*"

- *Covariance* between X_1 and X_2

$$\text{cov}(X_1, X_2) = E\Big[\{X_1 - E(X_1)\}\{X_2 - E(X_2)\}\Big]$$

- Will be > 0 if $X_1 > E(X_1)$ and $X_2 > E(X_2)$ or $X_1 < E(X_1)$ and $X_2 < E(X_2)$ tend to *happen together*
- Will be < 0 if $X_1 > E(X_1)$ and $X_2 < E(X_2)$ or $X_1 < E(X_1)$ and $X_2 > E(X_2)$ tend to *happen together*
- Will $= 0$ if X_1 and X_2 are $\underline{\perp\!\!\!\perp}$
- *Correlation* – covariance on a *unitless basis*

$$\rho_{X_1 X_2} = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)\text{var}(X_2)}}$$

- $-1 \leq \rho_{X_1 X_2} \leq 1$; $\rho_{X_1, X_2} = -1$ or 1 iff $X_1 = a + bX_2$

Conditional probability: Probabilistic statement of "*relatedness*"

- E.g., weight $Y > 200$ *more likely* for $X = 6$ than $X = 5$ feet tall
- X, Y *discrete*: conditional pmf *given* $X = x$ is function of y

$$f(y|x) = P(Y = y|X = x) = \frac{f(x, y)}{f_X(x)}, \quad f_X(x) > 0$$

  and satisfies $\sum_y f(y|x) = 1$ (a pmf for fixed x)

- X, Y *continuous*: conditional pdf *given* $X = x$ is function of y

$$f(y|x) = \frac{f(x, y)}{f_X(x)}, \quad f_X(x) > 0$$

  and satisfies $\int_{-\infty}^{\infty} f(y|x)\, dy = 1$ (a pdf for fixed x)

- Thus, the *conditional distribution* of Y given $X = x$ is *possibly different* for each x
- $Y|X$ denotes the *family* of probability distributions so defined

Conditional expectation: For $g(Y)$ a function of Y, define the *conditional expectation of Y given $X = x$*

$$E\{g(Y)|X = x\} = E\{g(Y)|x\} = \sum_y g(y)f(y|x) \quad \textit{discrete}$$

$$E\{g(Y)|X = x\} = E\{g(Y)|x\} = \int_{-\infty}^{\infty} g(y)f(y|x)\,dy \quad \textit{continuous}$$

- *Conditional expectation* is a function of x taking a value in $\Re$, *possibly different* for each x
- Thus, $E\{g(Y)|X\}$ is a *random variable* whose value depends on the value of X (and takes on values $E\{g(Y)|x\}$ as X takes on values x)
- *Conditional variance* defined similarly

Independence: If X and Y are *independent* random variables, then

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y)$$

and

$$E\{g(Y)|X = x\} = E\{g(Y)\} \quad \text{for any } x$$

so $E\{g(Y)|X\}$ is a *constant random variable* and equal to $E\{g(Y)\}$

Some probability distributions: "$\sim$" means "*distributed as*"

- $X \sim$ *Poisson*(λ) – a model for *counts* $x = 0, 1, 2, \ldots$

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad E(X) = \lambda, \ \text{var}(X) = \lambda$$

- *Normal* or *Gaussian* distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$. For $-\infty < x < \infty$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \ E(X) = \mu, \ \text{var}(X) = \sigma^2, \ \sigma > 0$$

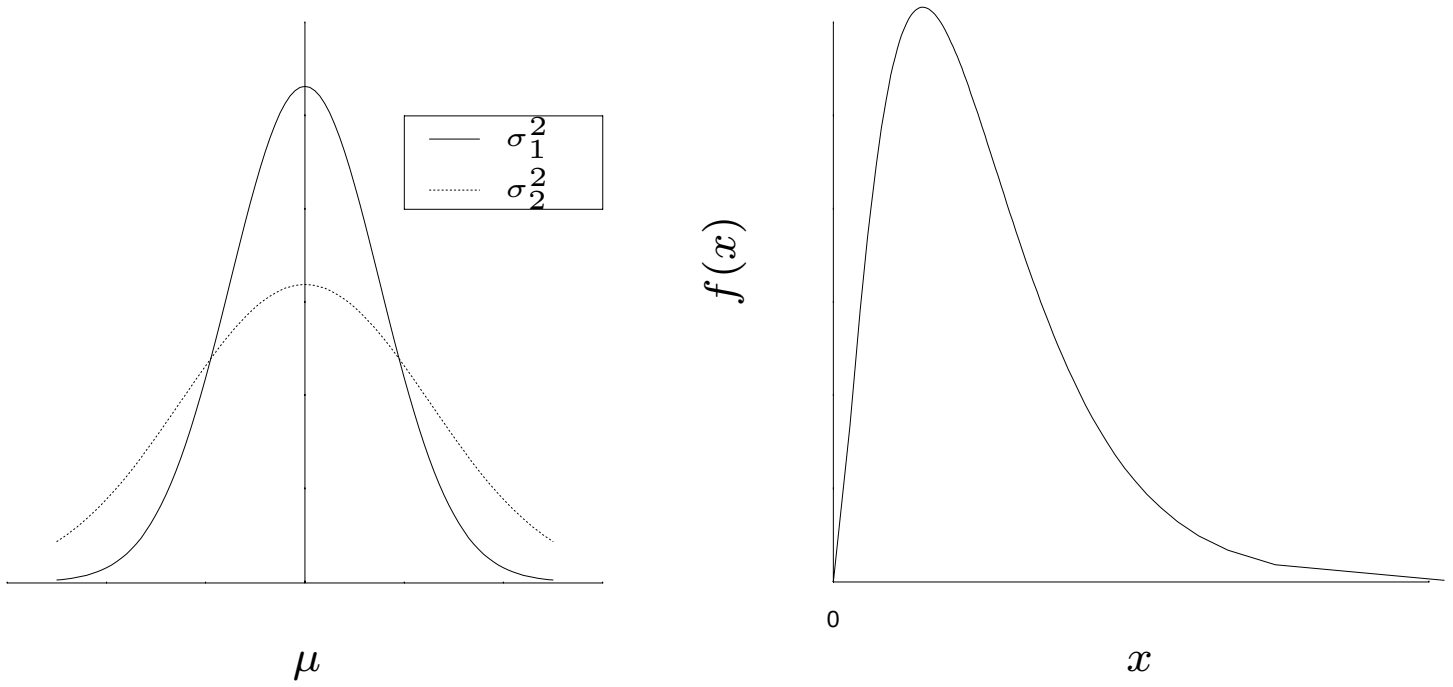
symmetric about μ, $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$ *standard normal*

- *Lognormal* distribution: $\log X \sim \mathcal{N}(\mu, \sigma^2)$

$$E(X) = e^{\mu+\sigma^2/2}, \ \text{var}(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2} \propto \{E(X)\}^2$$

Constant *coefficient of variation (CV)* $= \sqrt{\text{var}(X)}/E(X)$
("*noise-to-signal*") – does not depend on $E(X)$

(a) Normal pdfs (σ_1^2, σ_2^2) and (b) lognormal pdf:



(a) (b)

Multivariate normal distribution: *Random vector*

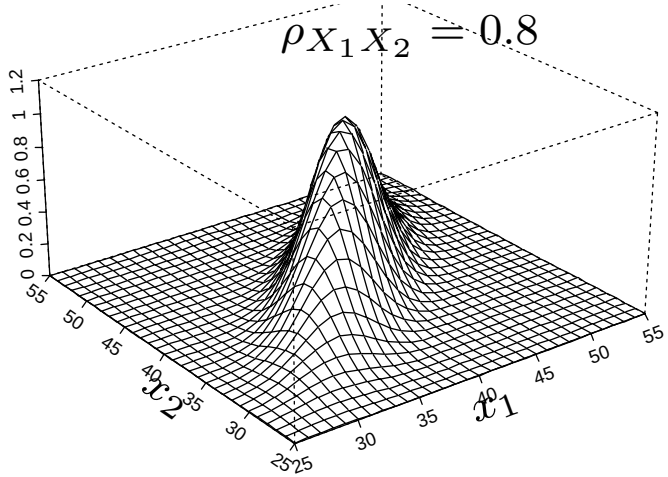
$X = (X_1, \ldots, X_p)^T$ has a *multivariate* (p-variate) *normal* distribution if $\alpha^T X \sim$ normal $\forall \alpha \in \Re^p$

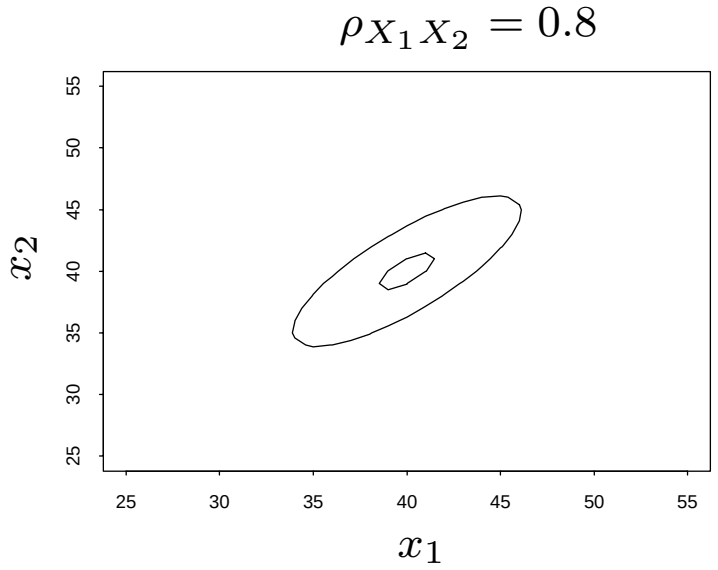
$$f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-(x-\mu)^T \Sigma^{-1} (x-\mu)/2\},$$

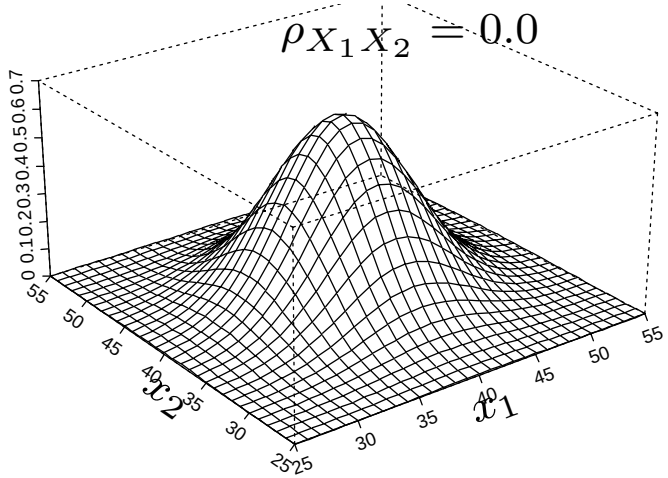
for $x = (x_1, \ldots, x_p)^T \in \Re^p$

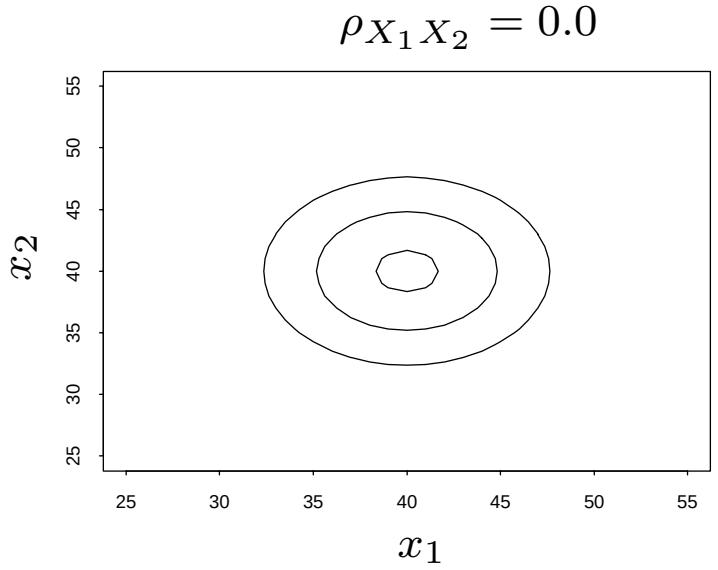
- $E(X) = \mu = (\mu_1, \ldots, \mu_p)^T = \{E(X_1), \ldots, E(X_p)\}^T$
- Σ $(p \times p)$ is such that $\Sigma_{jj} = \text{var}(X_j)$, $\Sigma_{jk} = \Sigma_{kj} = \text{cov}(X_j, X_k)$
- $\Sigma = E\{(X-\mu)(X-\mu)^T\}$ is the *covariance matrix*
- The *marginal* pdfs are *univariate* normal

Two bivariate ($p = 2$) normal pdfs:









3. Statistical models

Basic idea: *Random variables* and their *probability distributions* are the building blocks of *statistical models*

- A *statistical model* is a representation of the *mechanism* by which *data* are assumed to arise
- Phenomena that are *subject to variation* and hence give rise to *uncertainty* in the way data may "*turn out*" are represented by *random variables*
- Assumptions on the *probability distributions* for these random variables represent assumptions on the nature of such variation
- Return to the *theophylline example* for a demonstration...

Recall: For subject 12

$$Y_j = f(t_j, \theta) + \epsilon_j, \quad j = 1, \ldots, n$$

$$f(t, \theta) = \frac{k_a F D}{V(k_a - k_e)}\{e^{-k_e t} - e^{-k_a t}\}, \quad \theta = (k_a, k_e, V)^T$$

- Aggregate effects of *measurement error*, "*biological fluctuations*," other phenomena represented by *random variable* ϵ_j
- The *assumed probability distribution* of ϵ_j, and hence that of Y_j reflects assumed features of these phenomena

Aside: More formally, could observe the PK process at *any time* $\Rightarrow$

$$Y(t) = f(t, \theta) + \epsilon(t), \quad t \geq 0$$

- $Y(t)$ [and $\epsilon(t)$] is a *stochastic processes* with *sample paths* $y(t)$
- $Y_j = Y(t_j)$, $\epsilon_j = \epsilon(t_j)$

Example–Building a statistical model for subject 12 data:

Make *assumptions* that characterize beliefs about *variation*

$$Y_j = f(t_j, \theta) + \epsilon_j$$

Assumption 1: *Nature of* ϵ_j – "*additive effects*"

$$\epsilon_j = \epsilon_{1j} + \epsilon_{2j}$$

- ϵ_{1j} represents *measurement error* that could be committed at t_j
- ϵ_{2j} represents "*fluctuation*" that might occur at t
- *Continuous* random variables – concentrations *in principle* can take on *any value* (although we may be limited in what we may actually observe due to resolution of measurement)
- *Random vectors* $\epsilon_1 = (\epsilon_{11}, \ldots, \epsilon_{1n})^T$, $\epsilon_2 = (\epsilon_{21}, \ldots, \epsilon_{2n})^T$, $Y = (Y_1, \ldots, Y_n)^T$

Assumption 2–Measurement error: Some "*reasonable*" assumptions on aspects of the *joint probability distribution* of ϵ_1

- Measuring device is *unbiased* – does not systematically err in a particular direction $\Rightarrow$

$$E(\epsilon_{1j}) = 0 \quad \text{for each } j = 1 \ldots, n$$

(All possible errors for measuring concentration for the sample taken at any t_j "average out" to zero)

- In fact, negative or positive errors are *equally likely* $\Rightarrow$ the *marginal probability density* of ϵ_{1j} is *symmetric* for each j
- Measurement errors at any two times t_j, $t_{j'}$ are "*unrelated*"

$$\epsilon_{1j} \perp\!\!\!\perp \epsilon_{1j'} \quad \Rightarrow \quad \text{cov}(\epsilon_{1j}, \epsilon_{1j'}) = 0$$

- *Variation* among all errors that might occur at any t_j is the *same* $\Rightarrow$

$$\text{var}(\epsilon_{1j}) = \sigma_1^2$$

for all j (unaffected by time or "actual concentration" in the sample at t_j) – *is this realistic?*

Assumption 3–"Fluctuations": Some "*reasonable*" assumptions on aspects of the *joint probability distribution* of ϵ_2

- Fluctuations tend to "track" the smooth trajectory $f(t, \theta)$ over time (sample path) but can be "above" or "below" at any point in time $\Rightarrow$

$$E(\epsilon_{2j}) = 0$$

(All possible fluctuations at any particular time "average out" to zero)

- In fact, negative or positive fluctuations at a particular time are *equally likely* $\Rightarrow$ the *marginal probability density* of ϵ_{2j} is *symmetric*
- *Variation* among fluctuations that might occur at any t_j is *same* $\Rightarrow$

$$\text{var}(\epsilon_{2j}) = \sigma_2^2$$

- Fluctuations "*close together*" in time (at times t_j, $t_{j'}$) tend to behave "*similarly*," with extent of "*similarity*" decreasing as $|t_j - t_{j'}| \uparrow$

$$\text{cov}(\epsilon_{2j}, \epsilon_{2j'}) = C(|t_j - t_{j'}|) \Rightarrow \text{corr}(\epsilon_{2j}, \epsilon_{2j'}) = c(|t_j - t_{j'}|)$$

for decreasing functions $C(\,\cdot\,)$, $c(\,\cdot\,)$ with with $C(0) = \sigma_2^2$ and $c(0) = 1$

Assumption 3–"fluctuations," continued:

- E.g., for $\text{corr}(\epsilon_{2j}, \epsilon_{2j'}) = c(|t_j - t_{j'}|)$,

$$c(u) = \exp(-\phi u^2)$$

(so correlation between fluctuations at two times is *nonnegative*, reflecting "*similarity*")

- Extent and direction of measurement error at any time t_j *unrelated* to fluctuations at t_j or any other time $\Rightarrow$

$$\epsilon_{1j} \perp\!\!\!\perp \epsilon_{2j'}$$

for any t_j, $t_{j'}$, $j, j' = 1, \ldots, n$

Remarks:

- The foregoing assumptions are not the *only* assumptions one could make, but *exemplify* the considerations involved
- The *normal probability distribution* is a natural choice to represent the assumption of *symmetry*

Result: *Taken together*, these assumptions imply

$$Y \sim \mathcal{N}_n\{f(\theta), \sigma_1^2 I_n + \sigma_2^2 \Gamma\} \quad \psi = (\theta^T, \sigma_1^2, \sigma_2^2, \phi)^T \tag{1}$$

- $f(\theta) = \{f(t_1, \theta), \ldots, f(t_n, \theta)\}^T$
- I_n is $(n \times n)$ identity matrix, Γ is $(n \times n)$

$$\Gamma = \begin{pmatrix} 1 & e^{-\phi(t_1 - t_2)^2} & \cdots & e^{-\phi(t_1 - t_n)^2} \\ e^{-\phi(t_1 - t_2)^2} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & e^{-\phi(t_{n-1} - t_n)^2} \\ e^{-\phi(t_1 - t_n)^2} & \cdots & e^{-\phi(t_{n-1} - t_n)^2} & 1 \end{pmatrix}$$

- Each *marginal* is a normal density, e.g. $Y_j \sim \mathcal{N}\{f(t_j, \theta), \sigma_1^2 + \sigma_2^2\}$
- $E(Y_j) = f(t_j, \theta)$, $\text{var}(Y_j) = \sigma_1^2 + \sigma_2^2$, $\text{cov}(Y_j, Y_{j'}) = \sigma_2^2 e^{-\phi(t_j - t_{j'})^2}$
- *Interpretation* – $f(t, \theta)$ is the result of averaging across all possible *sample paths of the fluctuation process* and *measurement errors*, so representing the "*inherent trajectory*" for subject 12

Common simplification:

- If the t_j are *far apart in time*, $|t_j - t_{j'}|$ is *large*, and hence $\exp\{-\phi(t_j - t_{j'})^2\}$ *close to zero* $\Rightarrow$ "correlation among fluctuations at $t_1, \ldots, t_n$ is *negligible*"

- *Approximate* by assuming $\epsilon_{2j} \perp\!\!\!\perp \epsilon_{2j'} \Rightarrow \text{cov}(\epsilon_{2j}, \epsilon_{2j'}) = 0$ and thus $\Gamma = I_n \Rightarrow$

$$Y_j \perp\!\!\!\perp Y_{j'} \quad \Rightarrow \quad \text{cov}(Y_j, Y_{j'}) = 0,$$

  and $\text{var}(Y_j) = \sigma^2 = \sigma_1^2 + \sigma_2^2$

- The *statistical model* becomes

$$Y \sim \mathcal{N}_n\{f(\theta), \sigma^2 I_n\}, \quad \psi = (\theta^T, \sigma^2)^T \tag{2}$$

4. Statistical inference I

Key point: A *statistical model* like (1) or (2) describes *all possible probability distributions* for *random vector* Y representing the *data generating mechanism* for observations we might see at $t_1, \ldots, t_n$

- E.g., for (1), *possible probability distributions* are specified by different values of the *parameter* $\psi = (\theta^T, \sigma_1^2, \sigma_2^2, \phi)^T \in \Psi$
- *The big question*: Which value of ψ truly governs the mechanism?
- In particular, we are interested in θ ($\sigma_1^2, \sigma_2^2, \phi$ are required to describe things fully, but are a *nuisance* . . . more later)

Objective: If we *collect data* [so observe *a single realization* of $Y = (Y_1, \ldots, Y_n)^T$], what can we learn about ψ?

- . . . and how can we account for the fact that things could have turned out *differently* (i.e., a *different realization*)?

Conceptually:

- Think of the statistical model as a *formal representation* of the "*population*" of *all possible realizations* of $Y_1, \ldots, Y_n$ we would ever see
- When we collect data, we observe a *sample* from the *population*; i.e., a *single realization* of $Y_1, \ldots, Y_n$, $y_1, \ldots, y_n$

Objective, restated: What can we learn about the "*true value*" of ψ (which determines the nature of the *population*) from a *sample*?

- How *uncertain* will we be?

Statistical inference (loosely speaking): Making statements about a *population* on the basis of only a *sample*

Parameter (point) estimation: Construct a *function* of $Y_1, \ldots, Y_n$ that, if evaluated at a *particular realization* $y_1, \ldots, y_n$, yields a numerical value that gives information on the *true value* of ψ

- *Estimator:* The function itself
- *Estimate:* The numerical value for the particular realization
- *Estimation:* Used both to denote the procedure (*estimator*) and actual calculation of a numerical value (*estimate*)

Example: *Ordinary least squares (OLS)* estimator $\widehat{\theta}(Y)$

$$\arg\min_{\theta} \sum_{j=1}^{n} \{Y_j - f(t_j, \theta)\}^2$$

- For a *particular data set*, the OLS estimate $\widehat{\theta}$

$$\arg\min_{\theta} \sum_{j=1}^{n} \{y_j - f(t_j, \theta)\}^2$$

Remark: Distinction between *estimator* and *estimate* routinely *abused*

Convention: *Emphasis* that *estimator* is a *function* of Y usually *suppressed* (write $\widehat{\psi}(Y)$ as $\widehat{\psi}$)

Question: How "*good*" is $\widehat{\psi}(Y)$ as an *estimator*?

- A question about the *procedure*

Key idea: An *estimator* is a *function* of $Y \Rightarrow$ for any ψ, $\widehat{\psi}$ has a *probability distribution* (depending on that of Y and hence ψ)

- Each *realization* of Y yields a value – *sample space* for $\widehat{\psi}$
- "*All possible*" $\widehat{\psi}$ values from "*all possible data sets*," of which we *observe* only *one*
- *Large variance* – another realization might give *very different* result $\Rightarrow$ *lots of uncertainty*
- *Small variance* – similar answer from another realization $\Rightarrow$ *mild uncertainty*

Sampling distribution: The *probability distribution* of an *estimator*

- *Properties* characterize *uncertainty* in *estimation procedure* (*estimator*)
- *Unbiased estimator* – $E_\psi(\widehat{\psi}) = \psi$
- *Sampling covariance matrix* $\text{var}_\psi(\widehat{\psi})$
- *Sampling variance and standard error* for kth component

$$\text{var}_\psi(\widehat{\psi}_k), \quad \sqrt{\text{var}_\psi(\widehat{\psi}_k)}$$

- *Key factors* determining sampling variance
  - Variance of Y (may be *out of our control*)
  - Sample size n (often *under our control*)

(Very) simple example: $f(t, \theta) = \theta_1 + \theta_2 t$ in model $Y \sim \mathcal{N}_n\{f(\theta), \sigma^2 I_n)$, $\psi = (\theta^T, \sigma^2) \Rightarrow f(\theta) = X\theta$

- *OLS estimator* $\widehat{\theta}(Y) = (X^T X)^{-1} X^T Y$
- *Unbiased* – $E_\psi(\widehat{\theta}) = \theta$, *Sampling covariance*

$$\text{var}_\psi(\widehat{\theta}) = \sigma^2 (X^T X)^{-1} = \sigma^2 n^{-1} (n^{-1} X^T X)^{-1}$$

  - Depends on σ^2 and n, $\text{var}(\widehat{\theta}_k) = \sigma^2 \{(X^T X)^{-1}\}_{kk}$

- *Sampling distribution* – $\widehat{\theta} \overset{\psi}{\sim} \mathcal{N}_n\{\theta, \sigma^2 (X^T X)^{-1}\}$

Absolute necessity: Report of *estimate* should *always be accompanied* by *estimate* of *standard error*

$$\widehat{\sigma}^2 = (n-p)^{-1} \sum_{j=1}^{n} \{Y_j - f(t_j, \widehat{\theta})\}^2 \Rightarrow SE(\widehat{\theta}_k) = \sqrt{\widehat{\sigma}^2 \{(X^T X)^{-1}\}_{kk}}$$

- If $\widehat{\theta}_2 = 2.0$ and $SE(\widehat{\theta}_2) = 3.0 \Rightarrow$ *pretty uncertain*
- If $\widehat{\theta}_2 = 2.0$ and $SE(\widehat{\theta}_2) = 0.03 \Rightarrow$ *feeling pretty good!*

Confidence interval: *Probability statement* to *refine* statement of *uncertainty*

- *Linear example:* $T = \dfrac{\widehat{\theta}_k - \theta_k}{SE(\widehat{\theta}_k)} \overset{\psi}{\sim} t_{n-2}$ ("*Student's* t_{n-2} *dist'n*")
- If $t_{1-\alpha/2} \ni P(T \geq t_{1-\alpha/2}) = \alpha/2$

$$P\{\widehat{\theta}_2 - t_{1-\alpha/2}SE(\widehat{\theta}_2) \leq \theta_2 \leq \widehat{\theta}_2 + t_{1-\alpha/2}SE(\widehat{\theta}_2)\} = 1 - \alpha$$

- A probability pertaining to the *sampling distribution* of $\widehat{\theta}_2$: For "*all possible realizations of* Y *of size* n," probability is $1 - \alpha$ that *endpoints* of $[\,\widehat{\theta}_2 - t_{1-\alpha/2}SE(\widehat{\theta}_2)\ ,\ \widehat{\theta}_2 + t_{1-\alpha/2}SE(\widehat{\theta}_2)\,]$ include (the *fixed value*) θ_2
- Provides *more information* than just *SE*: how "large" or "small" SE must be *relative to* $\widehat{\theta}_2$ to feel "*confident*" that procedure of data generation and estimation provides a reliable understanding

Confidence interval: A probability statement about the *procedure* by which an *estimator* is constructed from a realization of Y

- *Interpretation:* "For all possible realizations of Y of size n, if we were to calculate the interval according to the *estimation procedure*, $(1-\alpha)\%$ of such intervals would '*cover*' θ_2"
- α is chosen by the *analyst*; e.g. $\alpha = 0.05$

Example:

- $\widehat{\theta}_2 = 2.0,\ SE(\widehat{\theta}_2) = 3.0$ gives $[-3.88, 7.88] \Rightarrow$ *no confidence*
- $\widehat{\theta}_2 = 2.0,\ SE(\widehat{\theta}_2) = 0.03$ gives $[1.94, 2.09] \Rightarrow$ *feeling pretty confident!*

Warning: The *numerical values themselves* are *meaningless* except for the *impression* they give about the *quality* of the *procedure*

- *Wrong:* The probability is $1-\alpha$ that θ_2 is between -3.88 and 7.88.
- *What we CAN say:* We are $(1-\alpha)\%$ "*confident*" that intervals constructed this way would "*cover*" θ_2

Parametric statistical model: A statistical model in which the *probability distribution* is *completely specified*, e.g.,

$$Y \sim \mathcal{N}_n\{f(\theta), \sigma^2 I_n\} \quad \text{for subject 12}$$

- $\Rightarrow$ *All* possible probability distributions are *multivariate normal* with this mean and covariance matrix (over all ψ)
- *If we believe this*, know *everything*

Semiparametric statistical model: A statistical model in which the *probability distribution* of Y is *only partially specified*, e.g.,

$$E(Y) = f(\theta), \quad \text{var}(Y) = \sigma^2 I_n$$

- *All* we're willing to say
- *Larger* class of *possible probability distributions*

Trade-off: *Fewer assumptions* $\Leftrightarrow$ *protection if wrong* $\Leftrightarrow$ could be "*too broad*"

Maximum likelihood estimation: Most *popular* approach for *parametric models*

The likelihood function: Suppose $p(y \mid \psi)$ is the pmf/pdf for the sample Y under the assumed *parametric model*

- The notation emphasizes that $p(y)$ is indexed by ψ (and is a *function* of y for *fixed* ψ)
- *Given that Y=y is observed*, the *likelihood function* of ψ defined by

$$L(\psi \mid y) = p(y \mid \psi)$$

- A *function of* ψ for *fixed* y

Discrete case: Suppose ψ_1 and ψ_2 are two possible values for ψ with

$$P_(Y = y|\psi_1) = L(\psi_1 \mid y) > L(\psi_2 \mid y) = P(Y = y|\psi_2)$$

- The y we *actually observed* is *more likely* to have occurred if $\psi = \psi_1$ than if $\psi = \psi_2 \Rightarrow \psi_1$ is "*more plausible*"

Maximum likelihood estimator (MLE): For each $y \in \mathcal{Y}$, let $\widehat{\psi}(y)$ be a parameter value where $L(\psi \,|\, y)$ attains its maximum as a *function* of ψ, with y *held fixed*. A *maximum likelihood estimator* for ψ is $\widehat{\psi}(Y)$

- *Intuitively reasonable* – MLE is the *parameter value* for which the observed y is "*most likely*"
- Has certain *optimality properties* under the *assumption* that the specified probability model is *correct*
  - *Invariance*
  - "*Most precise*" ($\approx$ smallest sampling variance for n "*large*")

Example: $Y \sim \mathcal{N}_n\{f(\theta), \sigma^2 I_n\}$

- *OLS estimator* is the *MLE*

In general: For *complex statistical models*, likelihood function is *complex* function of $\psi \Rightarrow$ computational issues

Remarks:

- *Identifiability of statistical models.* A *parameter* ψ for a family of probability distributions is *identifiable* if *distinct* values of ψ correspond to *distinct* pmf/pdfs. I.e., if the pmf/pdf for Y is $p(y \mid \psi)$ under the model, then

$$\psi \neq \psi' \Leftrightarrow p(y \mid \psi) \text{ is not the same function of } y \text{ as } p(y \mid \psi')$$

- *A feature of the model, NOT of an estimator or estimation procedure*
- *Difficulty with nonidentifiability in inference.* Observations governed by $p(y \mid \psi)$ and $p(y \mid \psi')$ look *the same* $\Rightarrow$ no way to know if ψ or ψ' is the true value (same likelihood function value)
- *How to fix nonidentifiability. Revise model*, impose *constraints*, make *assumptions*
- *Limitations of data.* Even if a model *is* identifiable, without sufficient data, it may be *practically impossible* to estimate some parameters, because the information needed is *not available* in the data

Sampling distribution for estimators in complex models:

- *Not possible* to derive a *closed form* expression for $\widehat{\psi}(Y) \Rightarrow$ derivation of *exact* sampling distribution *intractable*
- In *semiparametric models* – no *distribution assumption*
- *Approximation* – "*large sample* (*asymptotic*) *theory*" (such as $n \to \infty$)

For most "regular" estimators: $\widehat{\psi}_n$ $(p \times 1)$

- $\widehat{\psi}_n$ is *consistent* ("approaches" true ψ in probabilistic sense as $n \to \infty$)
- $\widehat{\psi}$ has "*approximate*" sampling distribution for $n \to \infty$

$$n^{1/2}(\widehat{\psi}_n - \psi) \dot{\sim} \mathcal{N}_p(0, C) \quad \text{or} \quad \widehat{\psi}_n \dot{\sim} \mathcal{N}_p(\psi, n^{-1}C_n), \; C = \lim_{n \to \infty} C_n$$

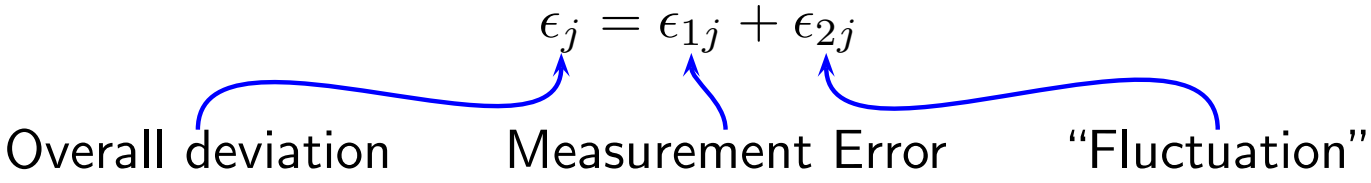
- *Comparison of competing estimators* – compare C, "*asymptotic relative efficiency*"

5. Modeling/inference: Independent data

Recall subject 12: Assumed *correlation* over time *negligible*

$$Y \sim \mathcal{N}_n\{f(\theta), \sigma^2 I_n\} \Rightarrow E(Y_j) = f(t_j, \theta), \ \text{var}(Y_j) = \sigma^2 \quad \underline{\|}$$

- *Assumed particular structure* $Y_j = f(t_j, \theta) + \epsilon_j$ with



  and $\sigma^2 = \sigma_1^2 + \sigma_2^2$

- OLS estimator is MLE estimator for θ *if we believe all of this*
- $\widehat{\sigma}^2 = (n-p)^{-1} \sum_{j=1}^{n} \{Y_j - f(t_j, \widehat{\theta})\}^2$
- *Do we believe*? *Can we check*?

$$\epsilon_j = Y_j - f(t_j, \theta)$$

$\Rightarrow$ *Plot* "*residuals*" $r_j = Y_j - f(t_j, \widehat{\theta})$ vs. "*predicted values*" $f(t_j, \widehat{\theta})$ or time t_j

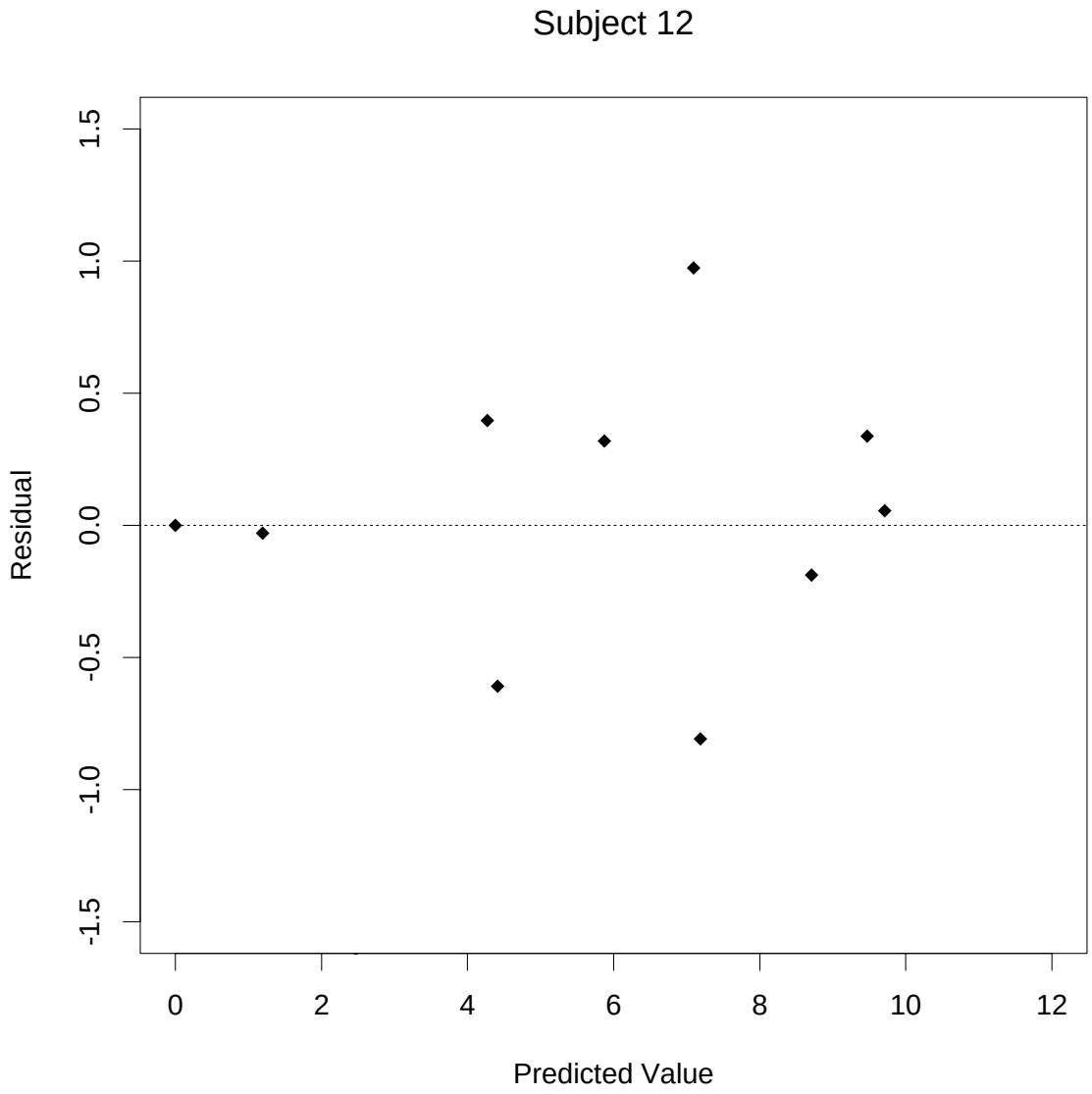
Informal impression:

- Magnitude of $r_j/\widehat{\sigma}$ *increases* with $f(t_j, \widehat{\theta})$
- Roughly *symmetric* about 0
- No apparent *pattern* with time

Suggestion: *Independence and normality* may be okay, but σ^2 *increases* with "*inherent*" concentration level

- "*Fluctuations*" and/or *measurement deviations* increase with inherent concentration

$r_j/\widehat{\sigma}$ **vs.** $f(t_j, \widehat{\theta})$**:**



Subject-matter knowledge – *Assay error* is dominant *source of variation* $\Rightarrow \quad \epsilon_j \approx \epsilon_{1j}$

- Error in determining concentrations *well-known* to be *greater* for *higher* concentrations $\Rightarrow$ suggests

$$\text{var}(\epsilon_j) = V\{f(t_j, \theta), \gamma\}, \quad V(u, \gamma) \uparrow \text{ in } u$$

- E.g., *popular choice of* V – $V\{f(t_j, \theta), \gamma\} = \sigma^2 f^{2\gamma}(t_j, \theta)$
- $\gamma = 1 \Rightarrow$ *constant CV* σ – "*Multiplicative error*"

$$Y_j = f(t_j, \theta)(1 + \delta_j), \; E(\delta_j) = 0, \; \text{var}(\delta_j) = \sigma^2, \quad \epsilon_j = f(t_j, \theta)\delta_j$$

$\Rightarrow Y_j \sim$ *normal*, *lognormal*, ...

- *Could include fluctuations* $\text{var}(\epsilon_j) = V(t_j, \gamma) + \sigma_2^2$

Result: $\psi = (\theta^T, \sigma^2, \gamma)^T$

- *Even though* θ is of central interest, *must* get the *entire* probability model *correct*

Inference: *MLE* or *something else*

- $Y_j \sim \mathcal{N}\Big[f(t_j, \theta), V\{f(t_j, \theta), \gamma\}\Big]$ – *MLE is NOT OLS* – OLS arises from *constant variance* (over j) assumption
- Using OLS *anyway* leads to *inefficient* inference on θ (and can be *biased* for *small* n)
- *Problem with MLE* – if *true* distribution is *NOT normal*, inference *compromised*; e.g., *sensitive* to *outliers*
- *Better approach* – *generalized least squares (GLS) estimator* (sort of like *weighted least squares*); solve

$$\sum_{j=1}^{n} V^{-1}\{f(t_j, \theta), \gamma\}\{Y_j - f(t_j, \theta)\}\frac{\partial}{\partial \theta} f(t_j, \theta) = 0$$

NOT the same as minimizing $\displaystyle\sum_{j=1}^{n} \frac{\{Y_j - f(t_j, \theta)\}^2}{V\{f(t_j, \theta), \gamma\}}$

(*Also* need estimator for $\gamma \ldots$)

Remarks:

- In *large samples*, *GLS estimator* is the "*best*" estimator if *only willing* to assume mean and variance (*semiparametric model*)
- If, in addition, *normality is correct*, *MLE* is "*better*," but can be led astray if *assumption of normality is wrong*, *GLS* is "*safer*"
- *Given* a means to calculate $f(t, \theta)$ at each t_j, *straightforward to implement* (including estimating γ)
- If assumption of *negligible correlation* is *incorrect*, all of this is suspect – need fancier techniques
- Note that must estimate *all components* of ψ, including those not of central interest – *efficiency of estimation* of θ will depend on that of estimators for "*nuisance parameters*"

Take-away messages:

- Must consider a *statistical model* that embodies *realistic assumptions*
- The *estimation procedure* is dictated by the *statistical model*
- Must estimate *all parameters*, not only those of interest
- OLS is *OFTEN NOT* the appropriate thing to do!

6. Hierarchical statistical models and inference

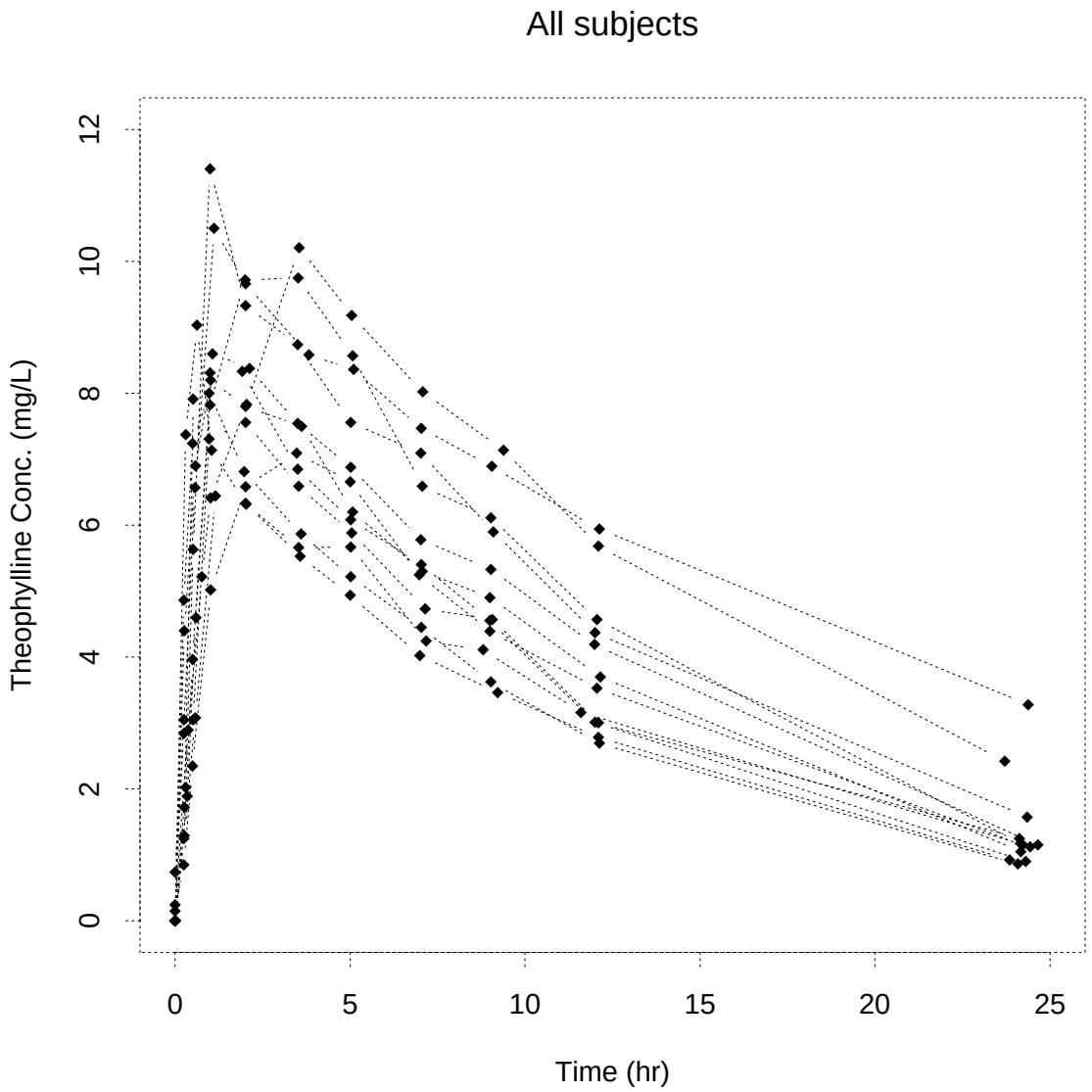
Recall objective 2: 12 subjects drawn from *population* of interest, $\theta = (k_a, k_e, V)^T$

- Learn about θ values in the *population* (*can we be more specific*?)
- Must acknowledge *measurement error*, etc, in data that would come from *each subject*
- Moreover, want to learn about *whole population* from a *sample* of only 12 subjects

Required: A *statistical model* that reflects the *data-generating mechanism*

- *Sample* m subjects from the *population*
- For the ith subject, ascertain concentration at *each of* n_i *time points* (could be *different* for each subject i)

All 12 subjects:



Formalization: "Learn about θ values in the population"

- *Conceptualize the* (infinitely-large) *population* as *all possible* θ (one for each subject) – how θs would "*turn out*" if we sampled subjects
- $\Rightarrow$ Represent by a (joint) *probability distribution* for θ (with *mean*, *covariance matrix*, etc)
- $\Rightarrow$ "*Average value* of θ" (*mean*), "*variability of* k_a, k_e, V" (diagonal elements of *covariance matrix*), "*associations between PK processes*" (off-diagonal elements)

Thus: Think of potential θ values if we draw m subjects *independently* as *independent random vectors* θ_i, each with *this probability distribution*, e.g.,

$$\theta_i \sim \mathcal{N}_p(\theta_*, D), \quad i = 1, \ldots, m$$

- When we *actually do this*, we obtain *realizations* of θ_i, $i = 1, \ldots, m$
- *Objective*, *formalized* – Estimate θ_* and D

Data for subject i: *Given* θ_i

- Concentrations arise from an *assumed mechanism* for an *individual subject* as before, e.g.

$$Y_{ij} = f(t_{ij}, \theta_i) + \epsilon_{ij}, \quad j = 1, \ldots, n_i$$

  $t_i = (t_{i1}, \ldots, t_{in_i})^T$ are the time points at which i would be observed, $f(t_i, \theta_i) = \{f(t_{i1}, \theta_i), \ldots, f(t_{in_i}, \theta_i)\}^T$

- $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$ *random vector* of observations from i
- *Same considerations* for ϵ_{ij} as before, e.g., $\epsilon_{ij} = \epsilon_{1ij} + \epsilon_{2ij}$

Result: Specify a family of *conditional probability distributions* for $Y_i|\theta_i$, e.g.,

$$Y_i|\theta_i \sim \mathcal{N}_{n_i}\{f(t_i, \theta_i), \sigma^2 I_{n_i}\}$$

- $E(Y_i|\theta_i) = f(t_i, \theta_i)$, $\text{var}(Y_i|\theta_i) = \sigma^2 I_{n_i}$
- Taking σ^2 *the same* for all i reflects belief that *assay error* is *similar* for blood samples from *any* subject

Or a fancier model:

$$Y_i|\theta_i \sim \mathcal{N}_{n_i}\{f(t_i,\theta_i), \sigma_1^2 R_i(\gamma) + \sigma_2^2 \Gamma_i(\phi)\}$$

$$R_i(\gamma) = \text{diag}\{f^{2\gamma}(t_{i1},\theta_i), \ldots, f^{2\gamma}(t_{in_i},\theta_i)\}, \quad (n_i \times n_i)$$

$$\Gamma_i(\phi)_{jj} = 1, \ \Gamma_i(\phi)_{jj'} = \exp\{-\phi(t_{ij} - t_{ij'})^2\}, \quad (n_i \times n_i)$$

- *Common* σ_1^2, σ_2^2, γ, ϕ across subjects reflects *assumption* of similar pattern of *variation* due to each source, could be modified
- So, e.g., $E(Y_{ij}|\theta_i) = f(t_{ij},\theta_i)$, $\text{var}(Y_{ij}|\theta_i) = \sigma_1^2 f^{2\gamma}(t_{ij},\theta_i) + \sigma_2^2$

Assumptions: $Y = (Y_{i1}^T, \ldots, Y_{in_i}^T)^T$ and $\theta = (\theta_1^T, \ldots, \theta_m^T)^T$

- Assume θ_i *independent* $\Rightarrow$ θ_i are *exchangeable* (all permutations of $\theta_1, \ldots, \theta_m$ have the *same joint probability distribution*)
- Also assume Y_i are $\perp\!\!\!\perp$ of each other, and $Y_i \perp\!\!\!\perp \theta_{i'}$, $i' \neq i$
- Thus, *joint pmf/pdf* of (Y^T, θ^T) is

$$p(y, \theta) = \prod_{i=1}^{m} p(y_i, \theta_i)$$

Hierarchical statistical model: *Combining* leads to a *two-stage hierarchy*

1. *Assumption* for *family of conditional probability distributions* for $Y_i|\theta_i \Rightarrow$ pmf/pdf $p(y_i|\theta_i)$ (n_i-dimensional, depending on t_i)
2. *Assumption* for *probability distribution* for $\theta_i \Rightarrow$ pmf/pdf $p(\theta_i)$ [same for all i; e.g., $\mathcal{N}_p(\theta_*, D)$]

$$p(y, \theta) = \prod_{i=1}^{m} p(y_i, \theta_i)$$

Remarks:

- Model in terms of *parameters* $\psi = \{\theta_*^T, \text{vech}(D)^T, \sigma_1^2, \sigma_2^2, \gamma, \phi\}^T$, $\text{vech}(D) =$ vector of distinct elements of D
- Note that $p(y_i|\theta_i)p(\theta_i) = p(y_i, \theta_i)$ for each i by *definition of conditional pmf/pdf* $\Rightarrow$

$$\prod_{i=1}^{m} p(y_i, \theta_i) = \prod_{i=1}^{m} p(y_i|\theta_i)p(\theta_i)$$

- The model contains *observable* and *unobservable* random components – Y_i, $i = 1, \ldots, m$, are observed, θ_i are *not* – *but both* are required in the formulation to reflect all *sources of variation*
- Because θ_i are *not observed*, would like the *probability distribution* of Y *alone*...

Result: (*Marginal*) *probability distribution* for *observable random vector* Y has pmf/pdf

$$\begin{aligned} p(y) &= \int p(y,\theta)\, d\theta \\ &= \int \prod_{i=1}^{m} p(y_i,\theta_i)\, d\theta_i \\ &= \prod_{i=1}^{m} \int p(y_i|\theta_i) p(\theta_i)\, d\theta_i \\ &= \prod_{i=1}^{m} \int p(y_i|\theta_i;\sigma_1^2,\sigma_2^2,\gamma,\phi) p(\theta_i;\theta_*,D)\, d\theta_i \qquad (1) \end{aligned}$$

where (1) highlights dependence on components of ψ

Objective: Find *estimator* for ψ

Natural approach: *Maximum likelihood*

$$L(\psi|y) = \prod_{i=1}^{m} \int p(y_i|\theta_i; \sigma_1^2, \sigma_2^2, \gamma, \phi) p(\theta_i; \theta_*, D)\, d\theta_i$$

- *MLE* maximizes $L(\psi|y)$ in ψ
- *Complication* – *intractable* p-dimensional *integration*
  - *Quadrature* or variant – bad in high dimensions
  - *Approximate integral* for large n_i
  - *Other approaches*. . .

Model refinement: θ_i *aren't exchangeable*

- *For example*, k_e is *associated with* weight, e.g., $E(k_{ei}|W_i = w_i) = \beta_1 + \beta_2 w_i$ for weight W_i
- *In general* $\theta_i|W_i \sim \mathcal{N}_p\{h(\beta, W_i), D\}$

Sampling distribution for MLE $\widehat{\psi}$: Usual approach

- *Approximate* for $m \to \infty$, n_i fixed
- *or* $m \to \infty$, $\min n_i \to \infty$

Software:

1. SAS `proc nlmixed`, macro `nlinmix`
   (`http://www.sas.com/rnd/app/da/new/danlmm.html`,
   `http://ftp.sas.com/techsup/download/stat/nlmm800.html`)
2. Splus `nlme()` function (`http://nlme.stat.wisc.edu/`)
3. `NONMEM`, SAS macro `nlmem`
   (`http://c255.ucsf.edu/nonmem0.html`,
   `http://www-personal.umich.edu/~agalecki/`)

Drawbacks – 1, 2 require user to provide *forward solution*; 3 have certain (but not arbitrary) systems hardwired, all can be flaky for high-dimensional θ

Other examples:

- (From Bedaux and Kooijman, 1994) – *Sample* of m insects treated to Cadmimum-contaminated food until equilibrium, then stopped – for Cd concentrations in post-food period (starting at $t = 0$)

$$\dot{C}(t) = -kC(t),\ Q(0) = Q_0, \quad \theta = (k, Q_0)^T$$

  for each insect

  - *Data collection* – at each of $0 \leq t_1 < \cdots < t_m$, a *single insect* is selected, Cd concentration measured (sacrifice the insect)
  - *Each insect* gives rise to a *single observation* Y_i $(n_i = 1)$
  - *Each insect* has its own $\theta = (k, Q_0)^T$ governing Cd kinetics $\Rightarrow$ insect i has $\theta_i = (k_i, Q_{0i})^T$
  - *Hierarchical model*

$$Y_i = f(t_i, \theta_i) + \epsilon_i \Rightarrow p(y_i|\theta_i), \quad \theta_i \sim p(\theta_i)$$

  - *Observable quantities* – $Y = (Y_1, \ldots, Y_m)^T \Rightarrow p(y_i)$

Other examples:

- (From Bortz, 2002) – *In vitro* experiment involving m "*identical*" cultures of HIV retrovirus strain – interest in process of culture growth over time, complex 4-dim system

$$\dot{X}(t) = g\{t, X(t), \theta\}$$

  - *Data collection* – at each of $0 \leq t_1 < \cdots < t_m$, a dish is chosen (at random) and total number of cells measured (destroy the dish)
  - *Although ideally*, all dishes should give rise to *identical* growth process, subtle *variation* in *conditions* for each dish $\Rightarrow$ different processes across dishes
  - $\Rightarrow$ *Each dish* has its own θ_i
  - Same *hierarchical model* as for insects

7. Statistical inference II

So far: We have considered the *classical*, *frequentist* approach to *statistical inference*

- *Objective* – *Estimate* parameters in a suitable statistical model, where the parameters are regarded as *fixed quantities*
- *Theme* – think of what would happen across *repeated samples* from the probability distribution dictated by the model ("*all possible samples we could have ended up with*")
- $\Rightarrow$ The *likelihood function* and *sampling distribution* are the cornerstones

Not surprisingly: As in many disciplines, there is a *competing* point of view and *philosophical debate*...

Bayesian inference: Basic idea

- Treat *parameters* like ψ as *random vectors*
- Make *inference* on ψ in terms of *probability statements* about ψ

Fundamental elements: In the Bayesian approach

- We *still specify* a probability distribution describing the "*data generating mechanism*" $\Rightarrow p(y|\psi)$, now viewed as a *conditional pmf/pdf*
- *Also* specify a pmf/pdf for the *prior distribution* of $\psi \Rightarrow p(\psi)$
- *Statistical inference* is based on implied *posterior distribution* with pmf/pdf $p(\psi|y) \Rightarrow$ the *conditional pmf/pdf* for ψ *given the data*
- The *mode* of this distribution is used as the "*estimate*"
- "*Uncertainty*" about this estimate is characterized by the entire *posterior distribution* (e.g., its variance)

Rationale and advantages:

- Can *formally incorporate* prior belief/opinion, knowledge, or info on ψ through specification of the *prior distribution*
- Can use the *prior distribution* to impose *constraints* on plausible values for ψ
- Interpretation can be *easier* than for some frequentist constructs (e.g., *confidence intervals*)

Possible disadvantages:

- *Sensitivity* of inferences to *choice of prior*?
- "*Who cares what you believe/think*?" (frequentist criticism)

Bayes' theorem: Cornerstone of Bayesian inference

$$p(\psi \,|\, y) = \frac{p(y \,|\, \psi)p(\psi)}{p(y)} = \frac{p(y \,|\, \psi)p(\psi)}{\int p(y \,|\, \psi)p(\psi)\, d\psi}$$

so that the *posterior pmf/pdf* satisfies $p(\psi \,|\, y) \propto p(y \,|\, \psi)p(\psi)$

- If ψ is a *scalar*, use Bayes theorem directly
- If ψ is *multidimensional*, use the *marginal posterior* pmf/pdf
- For example, $\psi = (\psi_1^T, \psi_2^T)$, $\psi_1 =$ parameter of *interest* (e.g., θ), $\psi_2 =$ "*nuisance parameters*"

$$p(\psi_1, \psi_2 \,|\, y) \propto p(y \,|\, \psi_1, \psi_2)p(\psi_1, \psi_2)$$

and the *marginal posterior* for ψ_1 is found by *averaging over* ψ_1

$$\begin{aligned} p(\psi_1 \,|\, y) &= \int p(\psi_1, \psi_2 \,|\, y)\, d\psi_2 \\ &= \int p(\psi_1 \,|\, \psi_2, y)p(\psi_2 \,|\, y)\, d\psi_2 \end{aligned} \quad \Rightarrow p(\psi_1 \,|\, y) \text{ is a } \textit{mixture}$$

$$p(\psi_1 \mid y) = \int p(\psi_1, \psi_2 \mid y)\, d\psi_2$$

Issue: Finding the *marginal posterior* involves potentially *high-dimensional integration* (more in a moment. . .)

Data generating model/likelihood specification: Same as before

Prior specification: Must choose a *probability distribution* and *values of its parameters* that reflect prior belief/knowledge/information about components of ψ

- *Prior elicitation* from "*experts*"
- Description of *historical data*, *results from literature*
- Sometimes criticized for focus on (analytical or computational) *convenience*

Simple example: $Y_j \overset{\parallel}{\sim} \mathcal{N}(\theta_1 + \theta_2 t_j, \sigma^2)$, $Y = (Y_1, \ldots, Y_n)^T \Rightarrow$

$$Y \sim \mathcal{N}_n(X\theta, \sigma^2 I_n) \text{ so that } p(y \mid \psi) \text{ is a normal pdf, } \psi = \theta$$

(σ^2 known)

- Choose *prior* $\psi \sim \mathcal{N}_3(T\psi^*, G) \Rightarrow p(\psi)$ is a normal pdf
- *Posterior pdf* $p(\psi \mid y)$ may be shown analytically to be that of

$$\mathcal{N}\{A(\sigma^{-2}X^T y + G^{-1}T\psi^*), A\}, \quad A = (\,\sigma^{-2}X^T X + G^{-1}\,)^{-1}$$

- *Posterior mode* (= posterior mean due to *symmetry*) is

$$A(\sigma^{-2}X^T y + G^{-1}T\psi^*)$$

  depends on *actual data observed*, y, and ψ^*, G characterizing *prior*

- Would need to specify ψ^* and G

Simple example, continued: *In fact*, if we take $G^{-1} = 0 \Rightarrow$ "*noninformative prior*," posterior mode becomes

$$(X^T X)^{-1} X y = \widehat{\theta}_{OLS}$$

and posterior distribution is

$$\mathcal{N}\{\widehat{\theta}_{OLS}, \sigma^2 (X^T X)^{-1}\}$$

- *Same estimate* as using *frequentist* inference
- Compare to *frequentist sampling distribution*

$$\widehat{\theta}_{OLS} | \theta \sim \mathcal{N}\{\theta, \sigma^2 (X^T X)^{-1}\}$$

Of course: Most problems are not this nice and yield different approach from frequentist

- Nuisance parameters, intractable integrals, etc
- *Luckily*, there is a *natural computational strategy*...

"Bayesian confidence intervals:" *Credible* interval or set C

- For chosen α, C satisfies

$$1 - \alpha = p(C|y) = \int_C p(\psi \,|\, y)\, d\psi$$

- "*The probability that ψ is in C, given the observed data y, is* $1 - \alpha$"

Reconciling frequentist and Bayesian approaches:

- Bayesian approach often leads to procedures with "*good*" properties in the frequentist sense (provided that prior distributions introduce only weak information)
- *In fact*, lead to *same* procedure for some statistical models!
- Under these conditions, *qualitatively similar* inferences
- Adopting a Bayesian perspective allows use of *computational strategies* in problems that would be intractable using frequentist tools (e.g., *hierarchical models* with *high-dimensional* integration)
- Bayesian approach provides *natural way* to impose *constraints*, exploit *previous knowledge*

8. Hierarchical models and Bayesian inference

Recall: *Theophylline example* $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$, $i = 1, \ldots, m$

$$p(y_i \,|\, \theta_i, \sigma^2): \quad Y_{ij} = f(t_{ij}, \theta_i) + \epsilon_{ij} \Rightarrow Y_i|\theta_i \sim \mathcal{N}_n\{f(t_i, \theta_i), \sigma^2 I_{n_i}\}$$

$$p(\theta_i \,|\, \theta_*, D): \quad \theta_i \sim \mathcal{N}_p(\theta_*, D)$$

- σ^2, θ_*, vech(D) are now *random variables/vectors*
- $p(\theta_i \,|\, \theta_*, D)$ is sometimes called the *prior*
- $\psi = \{\theta_*^T, \text{vech}(D)^T, \sigma^2\}^T$
- *Interested in* θ_*, D

To be Bayesian: Need to specify a *joint prior distribution* for ψ

- Often called the *hyperprior* in this context, with *hyperparameters* given by the analyst

Popular specification: Take θ_*, D, σ^2 *independent* with

$$\theta_* \sim \mathcal{N}_p(\delta, G), \quad D^{-1} \sim \text{Wishart}\{(\rho D_*)^{-1}, \rho\}, \quad \sigma^{-2} \sim \text{Gamma}(\nu/2, \nu\tau/2)$$

- *Hyperparameters* δ, G, ρ, D_*, ν, τ

Posterior distribution for θ_*: High-dimensional integration

$$p(\theta_*|y) = \frac{\int\int \left\{\prod_{i=1}^{m} \int p(y_i|\theta_i, \sigma^2) p(\theta_i|\theta_*, D)\, d\theta_i\right\} p(\theta_*|\delta, G) p(D|\rho, D_*) p(\sigma^2|\nu, \tau)\, dD d\sigma^2}{\int\int\int \left\{\prod_{i=1}^{m} \int p(y_i|\theta_i, \sigma^2) p(\theta_i|\theta_*, D)\, d\theta_i\right\} p(\theta_*|\delta, G) p(D|\rho, D_*) p(\sigma^2|\nu, \tau)\, dD d\sigma^2 d\theta_*}$$

- *Yuck*

Fortunately: It is possible to "do" these integrals by *simulation*, and end up with a *sample* from the desired *marginal posterior*, from which it may be approximated

- *Markov chain Monte Carlo* (MCMC) techniques

Basic idea: *Gibbs sampling*

- For random variables $U = (U_1, \ldots, U_K)$, given Y, the *full conditional distributions* $p_k(u_k | u_\ell \neq u_k, y)$ *completely determine* the *joint dist'n* $p(u_1, \ldots, u_K | y)$ and hence the *marginals* $p_k(u_k|y)$
- *Algorithm:* Start with $u_1^{(0)}, \ldots, u_K^{(0)}$

$$\begin{aligned}
\text{Draw } U_1^{(1)} &\sim p_1(u_1 | u_2^{(0)}, \ldots, u_K^{(0)}, y) \\
\text{Draw } U_2^{(1)} &\sim p_2(u_2 | u_1^{(1)}, u_3^{(0)}, \ldots, u_K^{(0)}, y) \\
&\vdots \\
\text{Draw } U_K^{(1)} &\sim p_K(u_K | u_1^{(1)}, \ldots, u_{K-1}^{(1)}, y)
\end{aligned}$$

- *Usefulness* – Can show $(U_1^{(t)}, \ldots, U_K^{(t)}) \dot{\sim} p(u_1, \ldots, u_K | y)$ as $t \to \infty$ $\Rightarrow$ can generate samples from the *joint posterior* (and hence get samples from *marginal posteriors*) by sampling from *full conditionals*!
- *For complex models*, actually need fancier algorithm – *FULL DETAILS ON MONDAY AFTERNOON!*

Full conditional distributions for the theophylline model:

$$(\theta_*|\sigma^2, D, \theta_i, i = 1, \ldots, m, y) \sim \mathcal{N}_p\{U(mD^{-1}\overline{\theta} + G^{-1}\delta), U\}$$

$$(D^{-1}|\sigma^2, \theta_*, \theta_i, i = 1, \ldots, m, y) \sim \text{ Wishart}\{(S + \rho D_*)^{-1}, m + \rho\}$$

$$(\sigma^{-2}|\sigma^2, D, \theta_*, \theta_i, i = 1, \ldots, m, y) \sim \text{ Gamma}\{(\nu + N)/2, (J + \nu\tau/2)\}$$

$$p(\theta_i|\sigma^2, D, \theta_*, \theta_\ell, \ell \neq i, y) \propto \exp(\sigma^{-2}J/2 - Q/2) \qquad (1)$$

- Except for (1), is *straightforward* to generate samples from these distributions
- To deal with (1), require *fancier approach* – *MONDAY AFTERNOON*

$U = (mD^{-1} + G^{-1})^{-1}$, $\overline{\theta} = m^{-1}\sum_{i=1}^m \theta_i$, $S = \sum_{i=1}^m (\theta_i - \theta_*)(\theta_i - \theta_*)^T$,
$J = \sum_{i=1}^m \{y_i - f(t_i, \theta_i)\}^T \{y_i - f(t_i, \theta_i)\}$, $Q = \sum_{i=1}^m (\theta_i - \theta_*)^T D^{-1}(\theta_i - \theta_*)$

Software:

1. `Bayesian inference Using Gibbs Sampling, BUGS` (`http://www.mrc-bsu.cam.ac.uk/bugs/`)
2. For common pharmacokinetic models `PKBUGS` (`http://www.med.ic.ac.uk/divisions/60/pkbugs_web/home.html`)
3. More general (e.g., physiologically-based PK models) `MCSim` (`http://fredomatic.free.fr/`)
4. These sites will lead you to more...

Drawbacks: 1 cannot do really complex models, 2 has only certain PK models hard-wired, 3 has steep learning curve

9. Closing remarks

- To take *appropriate account* of *sources of variation* in data, a deterministic model should be embedded in a *statistical model* that incorporates realistic assumptions
- *Inverse problem* should be regarded as *parameter estimation/inference* in the *statistical model framework*
- *Parameter estimation* should be *accompanied* by *assessment of uncertainty*
- *Personal belief* – *Bayesian formulation* implemented via *Markov chain Monte Carlo* methods are the way to go for *complex hierarchical models*
- *Important issue* for future work – *Design*

10. References and where to read more

Beduax, J.J.M. and Kooijman, S.A.L.M. (1994) Stochasticity in deterministic models. In *Handbook of Statistics, Vol. 12*, G.P. Patil and C.R. Rao, eds. Elsevier Science, pp. 561–581.

Bortz, D.M. (2002) Modeling, analysis, and estimation of an $in\ vitro$ HIV infection using functional differential equations. Ph.D. dissertation, Department of Mathematics, North Carolina State University.

Carlin, B.P. and Louis, T.A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Chapman and Hall/CRC.

Carroll, R.J. and Ruppert, D. (1988) *Transformation and Weighting in Regression*. Chapman and Hall.

Casella, G. and Berger, R.L. (2002) *Statistical Inference, Second Edition*. Duxbury Press.

Chen, M.H., Shao, Q.M., and Ibrahim, J.G. (2000) *Monte Carlo Methods in Bayesian Computation*. Springer.

Davidian, M. and Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall/CRC.

Gelman, A., Bois, F., and Jiang, J. (1996) Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* 91, 1400–1412.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995) *Bayesian Data Analysis*. Chapman and Hall/CRC.

Kaipio, J.P., Kolehmainen, V., Somersalo, E., and Vauhkonen, M. (2000) Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. *Inverse Problems* 16, 1487–1522.

Mosegaard, K. and Sambridge, M. (2002) Monte Carlo analysis of inverse problems. *Inverse Problems* 18, R29–R54.

Robert, C.P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. Springer.