

R Lab 1 - Simulating Longitudinal Data

Advanced Topics in Causal Inference

Assigned: September 07, 2021

Lab due: September 14, 2021 on bCourses. Please answer all questions and include relevant R code. You are encouraged to discuss the assignment in groups, but should not copy code or interpretations verbatim. Upload your own completed lab to bCourses.

Last lab:

1. Review of R programming and good programming practice.
2. Create basic function to generate data with $O = (W, A, Y)$ structure.
3. Review g-computation, IPTW, and TMLE estimators for point treatment intervention.

Goals for this lab:

1. Explore different data structures, inspired by real-world experiments.
2. Simulate data generating systems that give rise to data we observe.
3. Think of studies in which these data generating systems may occur.

Next lab:

We will translate causal questions into target causal parameters and intervene on the Structural Causal Models (SCMs) to evaluate them.

1 Introduction and Motivation

1.1 What is data simulation?

Data simulation is the process of sampling repeatedly from specified data generating distributions. In practice, this means generating random numbers from known distributions (that may or may not depend on each other), and repeating this process a large number of times. In a simulation, unlike reality, you know the truth about the underlying processes that give rise to your data.

1.2 Why simulate?

There are many reasons to simulate. Here are a few:

1. **By simulating, we get an understanding of the performance of estimators.** Suppose you come up with an estimator – a function that takes in as input your observed data and gives as output an estimate of a particular statistical estimand, say the g-computation formula corresponding (under assumptions) to the population average treatment effect. For a given sample size and data-generating process, how well will your estimator actually estimate your target statistical estimand? How biased will it be? How variable?

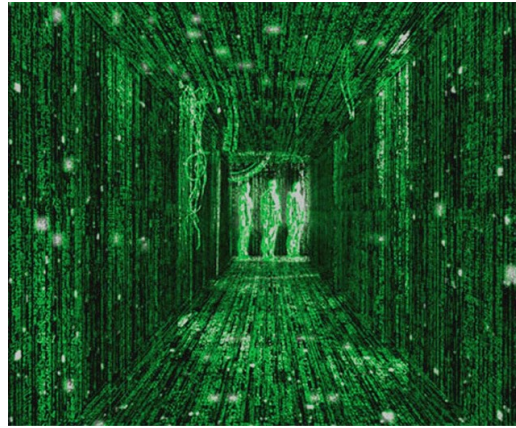


Figure 1: Simulation.

How is its behavior affected by things like data sparsity? What is its behavior like as the number of observations grows?

2. **Simulations are also helpful in evaluating your proposed approach to statistical inference.** Are your confidence intervals attaining their desired coverage? Across multiple repetitions of the same experiment, what percentage of the time will your proposed approach to confidence interval construction contain the true parameter value? Is your approach to hypothesis testing achieving its desired level of Type I error control? If the null hypothesis is true, what percentage of the time is your proposed approach to hypothesis testing end up rejecting the null?
3. **Simulations are useful for power analyses and sample size calculations.** You have a proposed study design and corresponding estimator. How big should your study be? For a given sample size and data generating process, how often will your estimator reject the null hypothesis? How big does your sample size need to be to ensure that the null hypothesis is rejected an acceptable proportion of the time?

2 This lab

2.1 Background

UC Berkeley School of Public Health is concerned about the amount of sleep students are getting. They've heard that lack of sleep is hurting students' academic performance and health. As part of a GSR project, you've been asked to study the effects of getting 8 or more hours of sleep on 1) students' statistics exam scores and 2) students' likelihood of getting sick.

Lucky for you, we have *perfect* knowledge of how sleep and other related variables work with each other to impact students' test performances and health outcomes. This means you'll know the generating system that gives rise to the variables you'd need to answer the scientific question of interest for your GSR.

For the first two data generating systems, you'll be asked questions related to the effect of 8 or more hours of sleep on students' statistics exam scores. For the last two data generating systems, we are interested in the effect of getting 8 or more hours of sleep on students' likelihood of becoming ill. In the next lab, we will evaluate these effects by intervening on the structural causal models presented below.

2.2 Link between observed data and causal model

In this lab we will simulate different data structures based on structural causal models (SCMs) that will help you answer your research questions on the effects of sleep. But what relates an SCM to the data we actually observe?

- In reality we might observe n copies of O , the observed ordered data structure of the above variables, from the observed data distribution, $P_0 \in \mathcal{M}$, (recall that \mathcal{M} is a statistical model of which the true distribution, P_0 , is an element).
- When working in the structural causal model framework, we assume that the observed data O were generated by sampling n times from a data generating system compatible with the causal model $\mathcal{M}^{\mathcal{F}}$. The distribution of the exogenous variables U together with the structural equations F identify the joint distribution of (U, X) , $P_{U,X}$, and thus the distribution of the observed data O (which is defined as a subset of X).
- In this lab we will define several processes for generating the data (U, X) , each of which defines a distribution $P_{U,X}$ of (U, X) , and by extension, a distribution P_0 of O . If we have background knowledge about a system we want to study, we can incorporate that when we specify the data generating process we wish to study in a simulation. In this lab, we will then use each data generating process that we define to generate a single sample of the observed data: n i.i.d. copies of O (in general, when using simulations to, for example, study estimator performance, one will repeat this process many times).
- Because we know the true underlying data generating process in a simulation, we can also calculate the true value of parameters of both statistical estimands and causal parameters under specific interventions on parameters of $P_{U,X}$ (next lab!).

2.3 To turn in:

For each of the 4 data structures listed below, answer the following questions:

1. **Write a function that generates n copies following the above data generating process.** The function should:
 - (a) Take in as input n , the number of observations you want to generate
 - (b) Output a dataframe with n i.i.d. copies of O .

Hint: Use the function in R Lab 0 that simulates a data-generating process for $O = (W, A, Y)$ as a template!
2. **Generate $n = 1000$ i.i.d. copies of O and store in a dataframe.**
3. **Show the first 6 lines and summary statistics of O** using the `head()` and `summary()` functions, respectively.
4. **Describe a study/experiment that could be represented by this data generating system.**

3 Data Structures

Data Structure 1: $O = (W, A, L, \Delta, \Delta Y)$

This data generating system represents an example with a baseline covariate, a single intervention, one covariate in between the intervention and outcome, and an outcome that is subject to missingness.

Let's define the following variables:

U - exogenous errors

W - baseline covariate that is a standardized measure of how many hours a student naps

A - the exposure of interest is sleep status for the night's sleep before the test. We define A as a binary variable indicating whether or not student got 8 or more hours of sleep the night before the test¹:

$$A = \begin{cases} 1, & \geq 8 \text{ hours of sleep the night before the test} \\ 0, & < 8 \text{ hours of sleep the night before the test} \end{cases}$$

L - a continuous covariate that measures the student's stress level the day of the test. The higher the value, the more the stress level.

Δ - a binary variable of missingness indicating whether or not we observe the outcome (e.g., whether or not the student turns in a statistics test). Specifically:

$$\Delta = \begin{cases} 1, & \text{if we observe the outcome} \\ 0, & \text{otherwise} \end{cases}$$

Y - student's statistics test score between 0 and 100.

Underlying data generating process, $P_{U,X}$:

Exogenous variables:

$$U_W \sim \text{Uniform}(\min = 0, \max = 1)$$

$$U_A \sim \text{Uniform}(\min = 0, \max = 1)$$

$$U_L \sim \text{Normal}(\mu = 2, \sigma^2 = 1^2)$$

$$U_\Delta \sim \text{Uniform}(\min = 0, \max = 1)$$

$$U_Y \sim \text{Normal}(\mu = 72, \sigma^2 = 0.3^2)$$

Structural equations F and endogenous variables:

$$W = U_W$$

$$A = \mathbb{I}[U_A < \text{expit}(0.01 * W)]$$

$$L = W + A + U_L$$

$$\Delta = \mathbb{I}[U_\Delta < \text{expit}(0.01 * (W + A + L))]$$

$$Y = L + 5 * A + 3 * W - 0.25 * A * W + U_Y$$

1. Set the seed to 252.

2. Write a function that simulates the specified data generating process. Call this function `generate_data1()`:

¹Can you think of some concerns or challenges that result from defining your exposure by categorizing an underlying continuous variable? There is an extensive discussion/debate in the literature about this. See for example:

- Petersen, Maya L. "Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs." *Epidemiology* 22.3 (2011): 378-381.
- Hernán, Miguel A., and Tyler J. VanderWeele. "Compound treatments and transportability of causal inference." *Epidemiology (Cambridge, Mass.)* 22.3 (2011): 368.

- (a) Create your exogenous variables, $U_W, U_A, U_L, U_\Delta, U_Y$, using the `rnorm()` and `runif()` functions. Careful with your parameters!
- (b) Recall that the *expit* function is the inverse of the logistic function:

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$$

$$\text{expit}(x) = \frac{1}{1 + e^{-x}}$$

In R, the *expit* function is called `plogis()`.

- (c) Use the `as.numeric()` function for endogenous indicator variables.
Example: $\mathbb{I}[X > 0]$ can be coded as `as.numeric(X>0)`.
- (d) For a given subject, if $\Delta = 0$, then Y is missing (which we code as `NA`). Otherwise, keep Y as is. You may use the `ifelse()` function on Y to code this.
Example: Suppose x is a vector with the numbers $[3, 1, 0, -1]$. If an element of x is greater than 0, then return 1, otherwise, return 0.

```
> x = c(3, 1, 0, -1)
> ifelse(x > 0, yes = 1, no = 0)
[1] 1 1 0 0
```

- (e) Make sure your function takes in an argument `n` (the number of copies of O) and returns back a dataframe with the endogenous variables.
3. **Generate $n = 1000$ copies and store in a dataframe.** Call this dataframe `ObsData1`.
 4. **Show the first 6 lines and summary statistics of O .**
 5. **Describe a study/experiment that could be represented by this data generating system.**
 6. **Bonus:** What does this data generating process encode about the effect of Δ on Y ? How would this be encoded in an SCM? Is it realistic in this setting?

Data Structure 2: $O = (\bar{L}(K), \bar{A}(K), Y)$ for $K = 4$
 $= (L(1), A(1), L(2), A(2), L(3), A(3), L(4), A(4), Y)$

This data structure represents a classic longitudinal experiment with time-varying treatment and covariates.

Variable definitions:

U - exogenous errors

$A(t)$ - the exposure of interest for this study is sleep status for the night's sleep starting on day t . In this example, we define $A(t)$ as a binary variable indicating whether or not student got 8 or more hours of sleep starting on day t :

$$A(t) = \begin{cases} 1, & \geq 8 \text{ hours of sleep starting on day } t \\ 0, & < 8 \text{ hours of sleep starting on day } t \end{cases}$$

$L(t)$ - a continuous covariate that measures the student's stress level on day t . Higher values correspond to more stress levels.

Y - the student's statistics test score between 0 and 100.

Underlying data generating process:

Exogenous variables:

$$U_{L(t)}, t = 1, \dots, 4 \sim \text{Normal}(\mu = 0, \sigma^2 = 1)$$

$$U_{A(t)}, t = 1, \dots, 4 \sim \text{Uniform}(\min = 0, \max = 1)$$

$$U_Y \sim \text{Normal}(\mu = 72, \sigma^2 = 3^2)$$

Structural equations F and endogenous variables:

$$L(1) = U_{L(1)}$$

$$A(1) = \mathbb{I}[U_{A(1)} < \text{expit}(0.001 * L(1))]$$

$$L(2) = A(1) + L(1) + U_{L(2)}$$

$$A(2) = \mathbb{I}[U_{A(2)} < \text{expit}(0.001 * (L(2) + A(1) + L(1)))]$$

$$L(3) = A(1) + L(1) + A(2) + L(2) + U_{L(3)}$$

$$A(3) = \mathbb{I}[U_{A(3)} < \text{expit}(0.001 * (L(1) + A(1) + L(2) + A(2) + L(3)))]$$

$$L(4) = A(1) + L(1) + A(2) + L(2) + A(3) + L(3) + U_{L(4)}$$

$$A(4) = \mathbb{I}[U_{A(4)} < \text{expit}(0.001 * (L(1) + A(1) + L(2) + A(2) + L(3) + A(3) + L(4)))]$$

$$Y = 0.3 * L(1) + A(1) + 0.5 * L(2) + A(2) + 0.7 * L(3) + A(3) + L(4) + A(4) - U_Y + 130$$

1. **Set the seed to 252.**
2. **Write a function `generate_data2()` that simulates the specified data generating process.** Use hints from the previous data structure to generate your endogenous and exogenous variables.
3. **Generate 1000 i.i.d. copies of O and store in a dataframe `ObsData2`.**
4. **Show the first 6 lines and summary statistics of O .**
5. **Describe a study/experiment that could be represented by this data generating system.**

Data Structure 3: $O = (L(1), A(1), Y(2), L(2), A(2), Y(3))$

Here, $Y(t)$ is an indicator variable describing whether or not the student has become ill by time t ; as such, it deterministically jumps to 1 and remains there once an individual has become ill. This is an example of a “survival” or “time to event” data structure. For outcomes such as this, once the event of interest has occurred, by definition, no future interventions can change it. We could thus truncate the observed data for an individual at the point that the event occurs (this is often done when defining survival data structures). However, in the longitudinal causal inference literature, for notational convenience, we often instead define the values of observed variables as something deterministic (such as last observed value) after the event occurs. This allows us to write down a data structure that extends through time $K + 1$ for all individuals (instead of a data structure where the number of time points varies by individual).

Definition of variables:

U - exogenous errors

$A(t)$ - the exposure of interest for this study is sleep status for the night’s sleep starting on day t . In this example, we define $A(t)$ as a binary variable indicating whether or not student got 8 or more hours of sleep starting on day t :

$$A(t) = \begin{cases} 1, & \geq 8 \text{ hours of sleep starting on day } t \\ 0, & < 8 \text{ hours of sleep starting on day } t \end{cases}$$

$L(t)$ - a continuous covariate that measures the student’s stress level on day t . Higher values correspond to higher stress levels.

$Y(t)$ is an indicator variable describing whether or not the student became ill by time t :

$$Y(t) = \begin{cases} 1, & \text{student became sick at or before time } t \\ 0, & \text{otherwise} \end{cases}$$

Underlying data generating process:

Exogenous variables:

$$U_{L(t)}, t = 1, 2 \sim \text{Normal}(\mu = 0, \sigma^2 = 1)$$

$$U_{A(t)}, t = 1, 2 \sim \text{Uniform}(\min = 0, \max = 1)$$

$$U_{Y(t+1)}, t = 1, 2 \sim \text{Uniform}(\min = 0, \max = 1)$$

Structural equations F and endogenous variables:

$$L(1) = U_{L(1)}$$

$$A(1) = \mathbb{I}[U_{A(1)} < \text{expit}(.001 * L(1))]$$

$$Y(2) = \mathbb{I}[U_{Y(2)} < \text{expit}(L(1) - 2 * A(1) - 6)]$$

$$L(2) = \begin{cases} A(1) + L(1) + U_{L(2)} & \text{if } Y(2) = 0 \\ \text{NA} & \text{if } Y(2) = 1 \end{cases}$$

$$A(2) = \begin{cases} \mathbb{I}[U_{A(2)} < \text{expit}(.001 * (L(1) + A(1) + L(2)))] & \text{if } Y(2) = 0 \\ \text{NA} & \text{if } Y(2) = 1 \end{cases}$$

$$Y(3) = \begin{cases} \mathbb{I}[U_{Y(3)} < \text{expit}(L(1) - 2 * A(1) + L(2) - A(2))] & \text{if } Y(2) = 0 \\ 1 & \text{if } Y(2) = 1 \end{cases}$$

1. **Set the seed to 252.**

2. **Write a function that simulates the data generating process.** Call the function `generate_data3()`.

(a) Create the exogenous variables as in the previous two data-generating processes.

(b) Create $L(1)$, $A(1)$ and $Y(2)$ as usual.

(c) If a subject has the event Y (i.e., gets sick) at time 2, then:

- The values that we assign deterministically to subsequent covariates and exposures will not affect the values of target causal parameters or their identification or corresponding estimands, nor will they be used in estimators of these estimands. We could assign them to have value equal to last observed value before the event occurred. However, for simplicity we here just set them to NA. That is: *if* $Y(2)$ is 1 for a given subject, *then* $L(2)$ and $A(2)$ should be NA. *Otherwise*, $L(2)$ and $A(2)$ should take on the value defined in the structural equations.

Hint: use the `ifelse()` function to generate $L(2)$ and $A(2)$'s missing values.

- Subjects should remain as having had the event at subsequent timepoints. So, *if* $Y(2)$ is 1 for a given subject, *then* $Y(3)$ should be 1, as well. *Otherwise*, $Y(3)$ should take on the value defined in the structural equations.

Hint: use the `ifelse()` function to generate $Y(3)$'s values.

3. **Generate 1000 copies of O and store in a dataframe called `ObsData3`.**

4. **Show the first 6 lines and summary statistics of O .**

5. **Describe a study/experiment that could be represented by this data generating system.**

Data Structure 4: $O = (L(1), C(1), A(1), Y(2), L(2), C(2), A(2), Y(3))$

We now introduce an example of a data structure with both a time to event (survival) outcome and a right censoring variable. Define $C(t)$ as an indicator variable whether or not the student dropped the class before day t . Thus, once $C(t)$ has jumped to 1 it remains there deterministically. Further, in general after a right censoring event (in this case, dropping the class) occurs, no more data is observed on an individual. However, similar to the survival example above, for notational convenience, we typically arbitrarily define the values of observed variables after a censoring event occurs (e.g., as deterministically equal to their last observed value). As for the survival example, this allows us to avoid having data structures of variable lengths, depending on the timing of failure time and censoring events, and instead to define a single data structure of length $K + 1$ that applies to all individuals, regardless of if and when they are censored or experience the failure event of interest.

Variable definitions:

U - exogenous errors

$A(t)$ - sleep status for the night's sleep starting on day t . $A(t)$ is a binary variable indicating whether or not student got 8 or more hours of sleep starting on day t :

$$A(t) = \begin{cases} 1, & \geq 8 \text{ hours of sleep starting on day } t \\ 0, & < 8 \text{ hours of sleep starting on day } t \end{cases}$$

$L(t)$ - a continuous covariate that measures the student's stress level on day t . Higher values of L correspond to higher stress levels.

$C(t)$ - an indicator variable describing whether or not the student dropped the class before day t .

$Y(t)$ is an indicator variable describing whether or not the student became ill by time t :

$$Y(t) = \begin{cases} 1, & \text{student became sick at or before time } t \\ 0, & \text{otherwise} \end{cases}$$

Underlying data generating process:

Exogenous variables:

$$U_{L(t)}, t = 1, 2 \sim \text{Normal}(\mu = 0, \sigma^2 = 1^2)$$

$$U_{C(t)}, t = 1, 2 \sim \text{Uniform}(\min = 0, \max = 1)$$

$$U_{A(t)}, t = 1, 2 \sim \text{Uniform}(\min = 0, \max = 1)$$

$$U_{Y(t+1)}, t = 1, 2 \sim \text{Uniform}(\min = 0, \max = 1)$$

Structural equations F and endogenous variables:

$$L(1) = U_{L(1)}$$

$$C(1) = \mathbb{I}[U_{C(1)} < \text{expit}(.001 * L(1) - 2)]$$

$$A(1) = \begin{cases} \mathbb{I}[U_{A(1)} < \text{expit}(.001 * L(1))] & \text{if } C(1) = 0 \\ \text{NA} & \text{if } C(1) = 1 \end{cases}$$

$$Y(2) = \begin{cases} \mathbb{I}[U_{Y(2)} < \text{expit}(L(1) - 2 * A(1) - 6)] & \text{if } C(1) = 0 \\ \text{NA} & \text{if } C(1) = 1 \end{cases}$$

$$L(2) = \begin{cases} L(1) + A(1) + U_{L(2)} & \text{if } Y(2) = 0 \text{ and } C(1) = 0 \\ \text{NA} & \text{if } Y(2) = 1 \text{ or } C(1) = 1 \end{cases}$$

$$C(2) = \begin{cases} \mathbb{I}[U_{C(2)} < \text{expit}(.001 * (L(1) + A(1) + L(2)) - 2)] & \text{if } C(1) = 0 \text{ and } Y(2) = 0 \\ \text{NA} & \text{if } C(1) = 1 \text{ or } Y(2) = 1 \end{cases}$$

$$A(2) = \begin{cases} \mathbb{I}[U_{A(2)} < \text{expit}(0.001 * (L(1) + A(1) + L(2)))] & \text{if } Y(2) = 0 \text{ and } C(2) = 0 \\ \text{NA} & \text{if } Y(2) = 1 \text{ or } C(2) = 1 \end{cases}$$

$$Y(3) = \begin{cases} \mathbb{I}[U_{Y(3)} < \text{expit}(L(1) - 2 * A(1) + L(2) - A(2))] & \text{if } Y(2) = 0 \text{ and } C(2) = 0 \\ 1 & \text{if } Y(2) = 1 \\ \text{NA} & \text{if } Y(2) = 0 \text{ and } C(2) = 1 \end{cases}$$

1. **Set the seed to 252.**

2. **Write a function `generate_data4()` that simulates the specified data generating process.**

(a) Generate the endogenous variables, $L(1)$ and $C(1)$ as usual.

(b) If a subject is censored at time t , we set the values of most variables after censoring has occurred deterministically. We could assign them to have a value equal to the last observed value before the event occurred. However, for simplicity, here we just set them to NA. An important exception is $Y(t)$ – once the failure event has been observed to occur, it is known to be deterministically equal to 1 thereafter.

Hint: use the `ifelse()` function to generate missing values of variables after right censoring occurs.

(c) As in the previous data structure, if a subject has the event (i.e., gotten sick) at time t (i.e., $Y(t) = 1$), then:

- They should remain as having had the event at subsequent timepoints. So, *if* $Y(t)$ is 1 for a given subject, *then* $Y(t+1), Y(t+2), \dots, Y(K+1)$ should be 1, as well. *Otherwise*, $Y(t+1), Y(t+2), \dots, Y(K+1)$ should take on the value defined in the structural equations.

Hint: use the `ifelse()` function to generate $Y(t+1), Y(t+2), \dots, Y(K+1)$'s values.

- Again, as in the previous data structure, the values that we assign deterministically to subsequent covariates and exposures will not affect the values of target causal parameters or their identification or corresponding estimands, nor will they be used in estimators of these estimands. We could assign them to have value equal to last observed value before the event occurred. However, for simplicity we here just set them to NA. That is: *if* $Y(t)$ is 1 for a given subject, *then*

$A(t+1), \dots, A(K)$ should be NA. Otherwise, $A(t+1), \dots, A(K)$ should take on the value defined in the structural equations. Same goes for generating L 's after an event has occurred.

Hint: use the `ifelse()` function to generate $L(t+1), A(t+1), \dots, L(K), A(K)$'s missing values.

3. **Generate 1000 copies of O and store in a dataframe `ObsData4`.**
4. **Show the first 6 lines and summary statistics of `ObsData4`.**
5. **Describe a study/experiment that could be represented by this data generating system.**

4 For Your Project: Simulations

Think through the following questions and apply them to the dataset you will use for your final project. This exercise is not required for R Lab 1 submission.

1. Define your variables

- (a) We have not yet worked on defining specific target casual parameters of interested. However, thinking ahead, what variables are you interested in intervening on? What are your outcome variables? When were your variables measured?

2. Go through first step of roadmap (i.e., specify the causal model representing real background knowledge).

- (a) Draw a DAG to represent how you believe the variables relate to each other based on background knowledge. Do this for two timepoints.
- (b) Using those relationships drawn in the previous step, define your structural equations generically; in other words, don't assuming distributions or functional forms yet. Do you have any prior knowledge on the functional forms?

3. Make histograms for continuous variables and tables for binary/categorical variables.

- (a) What shapes do the distributions seem to take? Based on the shape, what known distribution do you think that variable's error term is drawn from?
- (b) If you've picked a distribution for an exogenous variable, how would you parameterize it?

4. Is there missingness in your dataset?

- (a) Which variables are subject to missingness? The covariates? treatment? outcome?
- (b) Is there censoring? In other words, are there people who, at a certain timepoint, have missing values for the rest of the study?
- (c) Based on *a priori* knowledge, do you expect missingness to depend on other variables? Which ones?

5. Come up with more specific structural equations that relate the endogenous variables based on the previous two questions.

6. Create a function to simulate your data and generate $n = 1000$ copies of your O .

- (a) Check the histograms and summary statistics of the variables in your simulated data and see how they match up to your real data. Over the course of the class you can refine your data generating structure to match the data at hand.

5 Optional Feedback

You may attach responses to these questions to your lab. Thank you in advance!

1. Did you catch any errors in this lab? If so, where?
2. What did you learn in this lab?
3. Do you think that this lab met the goals listed at the beginning?
4. What else would you have liked to review? What would have helped your understanding?
5. Any other feedback?