Introduction to Epidemiological and Biostatistical Thinking

UW Neurology Fellowship

Marlena Bannick

7/23/2020

PhD Student, University of Washington Dept. of Biostatistics
Researcher, Institute for Health Metrics and Evaluation

Goal

Introduce you to epidemiological thinking and key (bio)statistical concepts that you can use to critically interpret scientific studies in health and medicine.

Learning Objectives (1/2)

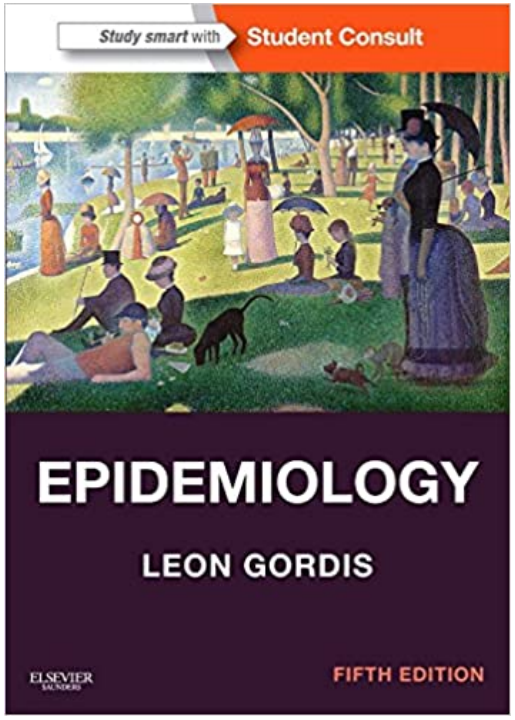
1. **Basics**. Identify key elements of an epidemiological study and how they relate to the scientific question
2. **Study Design**. Recognize the basic types of epidemiological study design and identify when each design is appropriate for the scientific question
3. **Bias**. Recognize sources of bias in study designs or measurements and understand how they might affect your ability to answer the scientific question

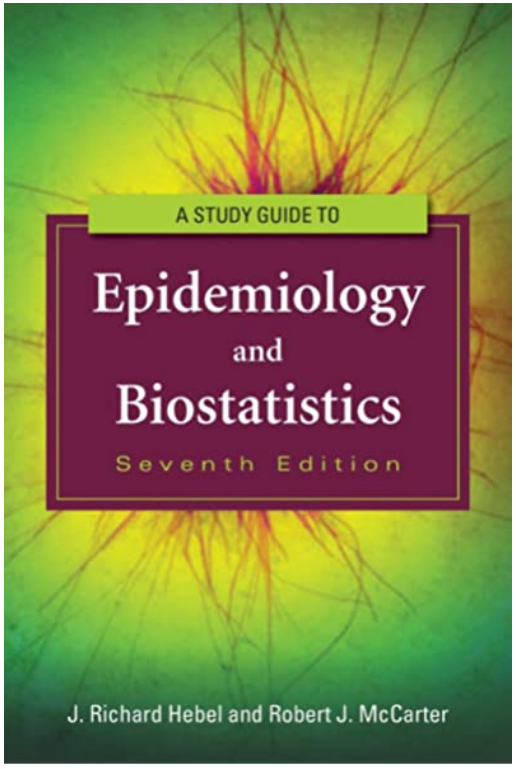
Learning Objectives (2/2)

4. **Modeling**. Understand how you can formulate your understanding about a data generating process, assumptions, and a hypothesis to test in a statistical model
5. **Inference**. Recognize the distinction between an effect size, a confidence interval, and a p-value as they relate to parameters that are estimated in a statistical model

Further Study

This lecture and follow-up discussion will be a very brief introduction, with some material borrowed from the following texts. These are good introduction texts to epidemiology and biostatistics:





Basics

A epidemiological study should be generated by a *scientific question of interest*. Broadly, you can think of these scientific questions falling into two main categories:

- **Descriptive**: What is the incidence rate of ischemic stroke (IS) in women aged 45 - 60 years old?
- **Inferential**: What is the effect of an experimental treatment on mortality following ischemic stroke in women aged 45 - 60?

From a statistical point of view it is not a clean distinction because you still use statistical tools to *infer* the incidence rate for a descriptive study.

Basics: Terminology

The questions *who, what, where, when* have never been more important than in the context of epidemiology!

Having a well-defined scientific question means having clear answers for the following components:

- **population**: Who is the group being studied?
- **exposure**: What is the group in study exposed to that you want to measure the effect of, and over what period of time?
- **outcome**: What outcome is being studied (either in relation to the exposure or on its own) and over what period of time?

The *why* is also important! Epidemiological studies should serve some purpose.

Basics: Measures

Once you've defined your target exposure, outcome, and population that makes up your scientific questions, understanding **measurement** of the outcomes is of utmost importance.

Some common outcome measurements in the context of health sciences are

- **prevalence**: proportion of a population with an outcome
- **incidence**: rate of getting the outcome among individuals in a population that did not already have the outcome ("risk")
- **remission**: rate of returning to be outcome-free among those that had the outcome

Think about denominators!

Basics: 2x2 Tables

With a binary exposure and a binary outcome, the results of a study will look something like this 2x2 table:

Table 1: Example 2x2 Table

	Outcome	No Outcome
Exposed	a	c
Unexposed	b	d

But there are *so many ways* to obtain that 2x2 table, so it is imperative to understand the study design behind the data!

Understanding study design will make it clear **what are the valid analyses** that can be performed on the data in that table.

Basics: Example

What are the exposure, outcome, and population for each of these scientific questions?

- **descriptive**: What is the incidence rate of ischemic stroke (IS) in women age 45 - 60 years old?
- **inferential**: What is the effect of an experimental treatment on mortality following ischemic stroke in women age 45 - 60?

Table 2: Basic Elements of Study Design

	Descriptive	Inferential
Exposure		
Outcome		
Population		

Basics: Example

What are the exposure, outcome, and population for each of these scientific questions?

- **descriptive**: What is the incidence rate of ischemic stroke (IS) in women age 45 - 60 years old?
- **inferential**: What is the effect of an experimental treatment on mortality following ischemic stroke in women age 45 - 60?

	Descriptive	Inferential
Exposure		experimental treatment
Outcome	ischemic stroke (IS)	death from IS
Population	women age 45-60 without IS	women age 45-60 with IS

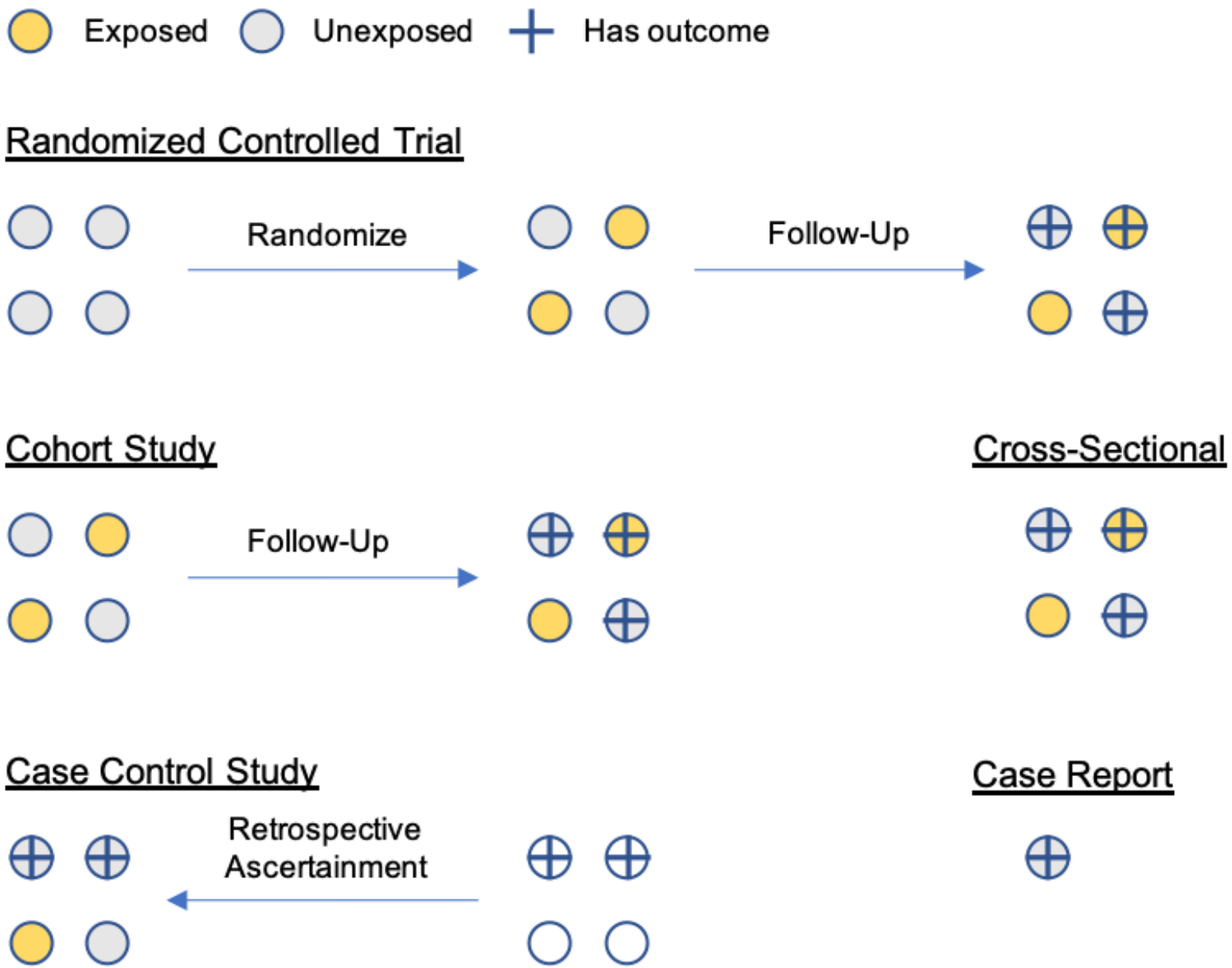
How would you make these questions more precise?

Study Design: Types

Starting with what is typically considered the studies that will provide the "strongest" evidence of a *causal* relationship between an exposure and an outcome:

- **Randomized controlled trials**: participants are *randomly* assigned to an exposure treatment or a control and followed up over time to record outcomes
- **Cohort studies**: participants are selected based on their exposure status and followed up to record outcomes
- **Case control studies**: participants are selected based on their outcome status and we inquire about exposure in the past
- **Cross-sectional studies**: measure exposure and outcome of participants at the same point in time (no temporal element)
- **Case reports**: report on the outcome status of one or a handful of interesting cases

Study Design: Diagram of Types



Study Design: Example

Recall our example: **What is the effect of an experimental treatment on mortality following ischemic stroke in women age 45 - 60?**

Explain how each of the following study types would be designed to answer this research question? Which is preferable?

- Randomized controlled trial
- Cohort study
- Case control study
- Cross-sectional
- Case report

Biases: Taxonomy

Biases in the epidemiological context are any factors in your study that **prevent you from being able to answer your precise scientific question**.

Biases may result from systematically flawed measurements of the outcome, the exposure, or the population, categorized generally as:

- **selection biases**: biases that are a function of the sampling or selection of participants for the study -> cannot generalize to target population
- **information biases**: biases that are a function of how the measurements on participants are taken

Some study designs may avoid certain types of bias, but it is crucial to always be on the lookout for sneaky biases when designing, analyzing, or reading a study.

Biases: Examples

Examples of biases include:

- **Loss to follow-up bias**: participants leave the study in such a way that it distorts the relationship between the exposure and outcome
- **Confounding bias**: the relationship between exposure and outcome among those in your study is *confounded* by other variables (more later)
- **Recall bias**: individuals are being asked about exposures or outcomes that they do not remember correctly
- **Social desirability bias**: individuals are not comfortable disclosing their true exposure or outcome status for fear of judgement by others

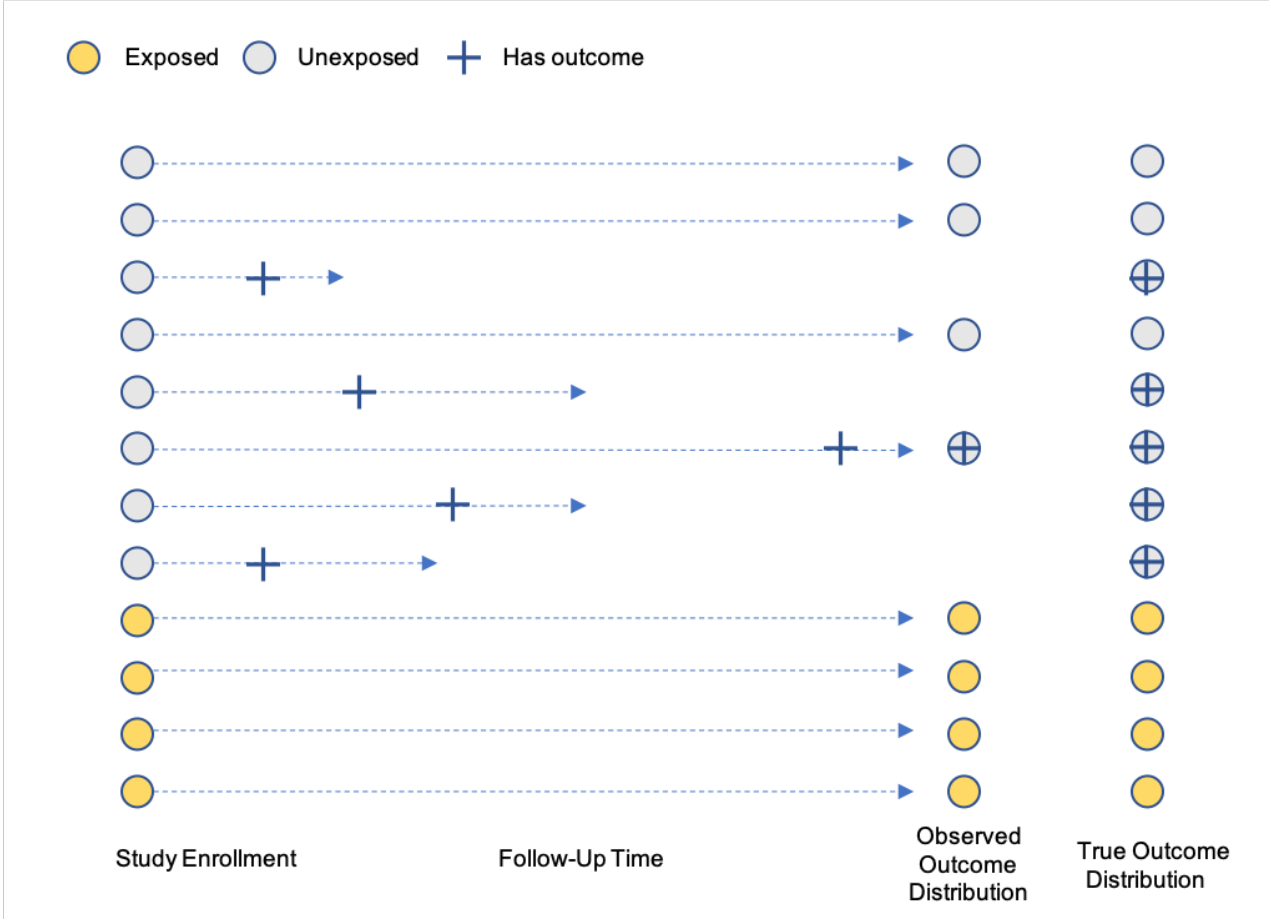
This is by no means an exhaustive list. See a catalogue of bias for a taxonomy and more examples.

Biases: Comments on Study Design

- Randomized controlled trials are designed to **eliminate bias**: statistically speaking, we do not expect there to be significant differences in the characteristics of the treatment groups
- Observational study designs like cohort studies and case control studies **observe what's already happening** – what if those that are exposed also have characteristics that make it more likely that they will have the outcome (**confounding**: more later)?
- Studies that rely on participants to self-ascertain, or to recall things from the past (e.g. case control studies) may result in systematic measurement error of exposure or outcome (information bias)

Biases: Selection Bias Diagram

An example of selection bias is loss to follow-up. Consider a situation where



Biases: Selection Bias Example

Recall our example inferential question: **What is the effect of an experimental treatment on mortality following ischemic stroke in women age 45 - 60?**

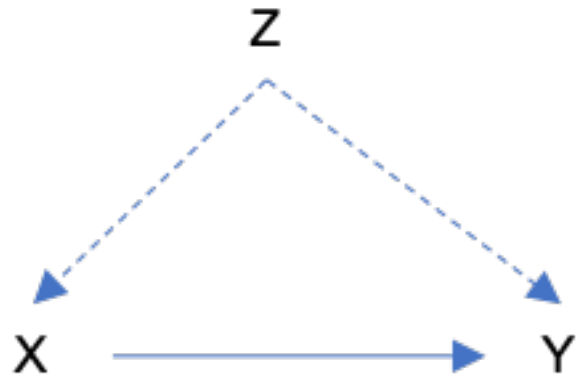
Consider the following sampling strategies:

1. Sample women aged 45 - 60 who have been discharged from the hospital following ischemic stroke, randomly assign some to experimental treatment.
2. Sample women aged 45 - 60 who have been admitted to the hospital for ischemic stroke, randomly assign some to experimental treatment.

Which may suffer from selection bias?

Biases: Confounding Diagram

Confounding occurs when there is a third, measured or unmeasured, factor Z that causes the exposure X and is associated (perhaps causally) with the outcome Y. If $X \rightarrow Z$, then it is not a confounder because Z is *in the causal pathway* between X and Y.

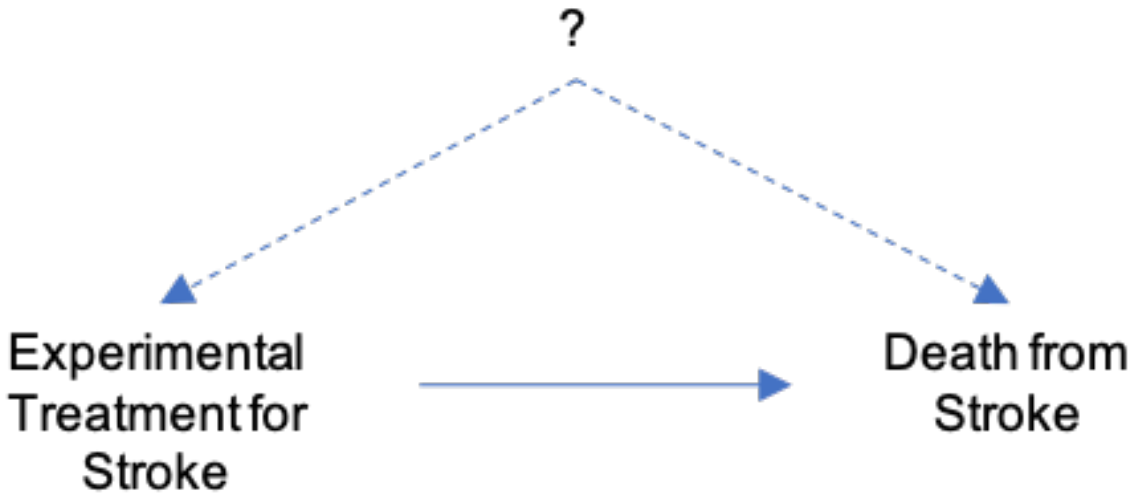


If we have **measured** Z, then there is hope that we can recover the true relationship between X and Y. If Z is unmeasured (which is often the case), it is much more difficult. In the case where Z is measured, there are standard techniques to "control" for Z.

Biases: Confounding Example

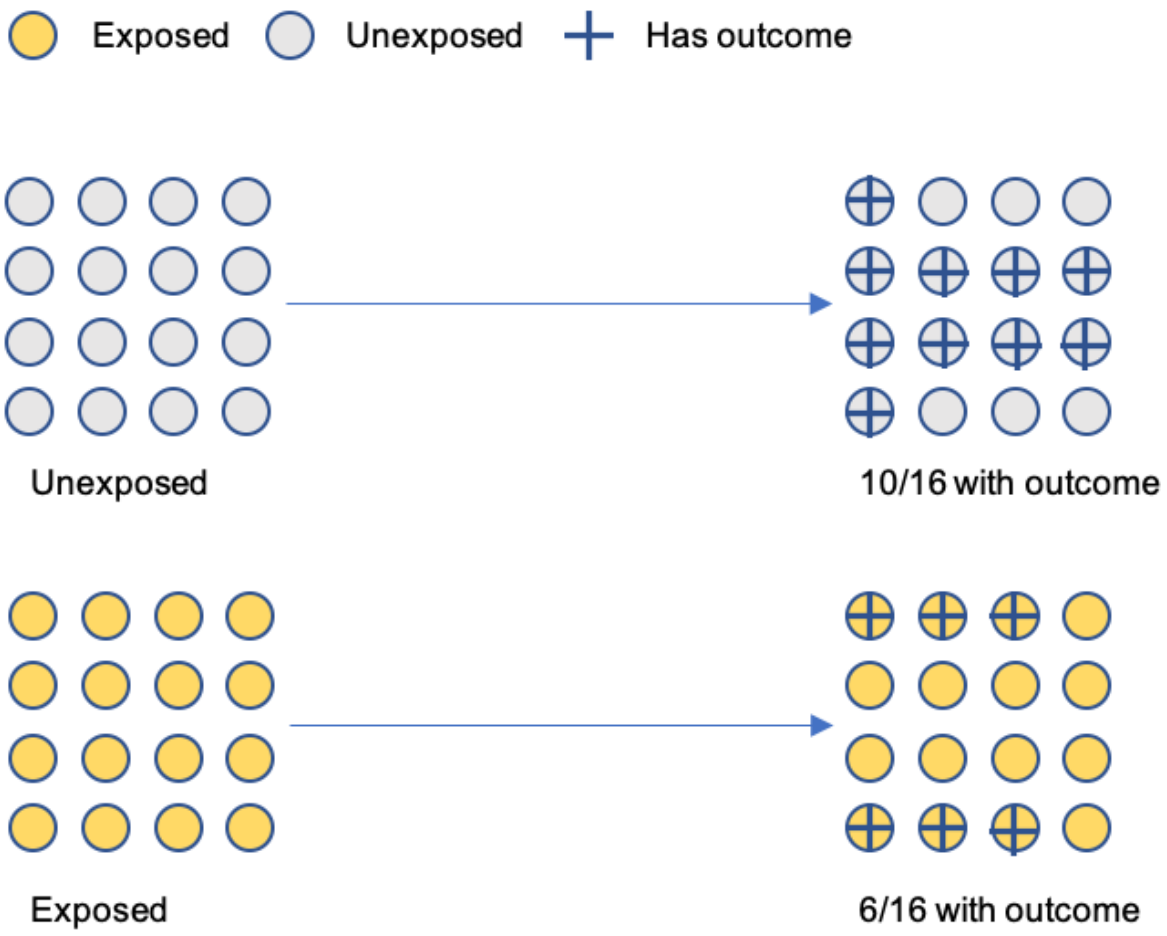
Again, recall our example inferential question: **What is the effect of an experimental treatment on mortality following ischemic stroke in women age 45 - 60?**

What if we do not assign the experimental treatment, but the physician decides whether or not to administer treatment to the patient?



Biases: Confounding and Stratification (1/5)

Consider a study where we have an equal number of exposed and unexposed participants, and we want to follow them over time to observe an outcome.

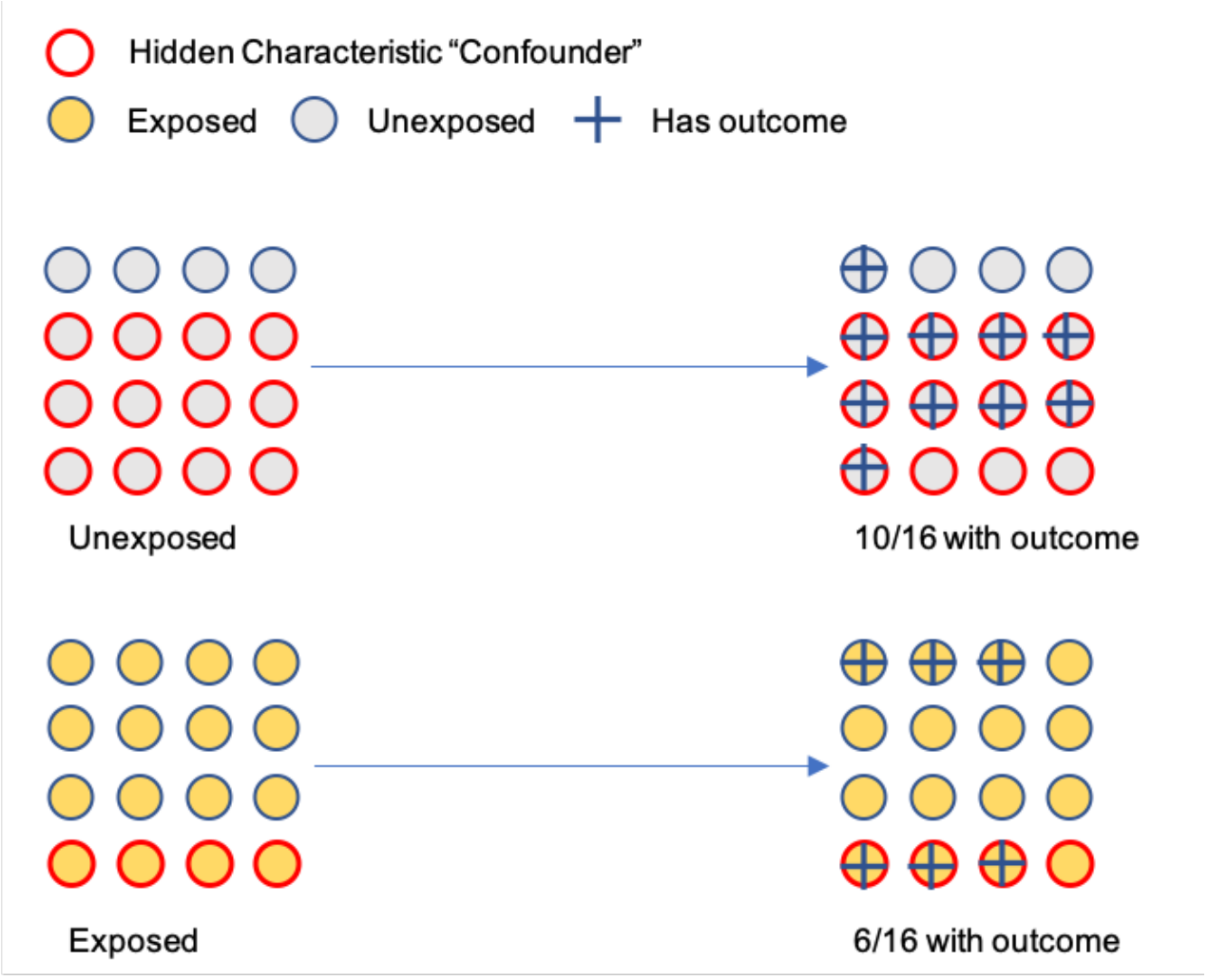


Biases: Confounding and Stratification (2/5)

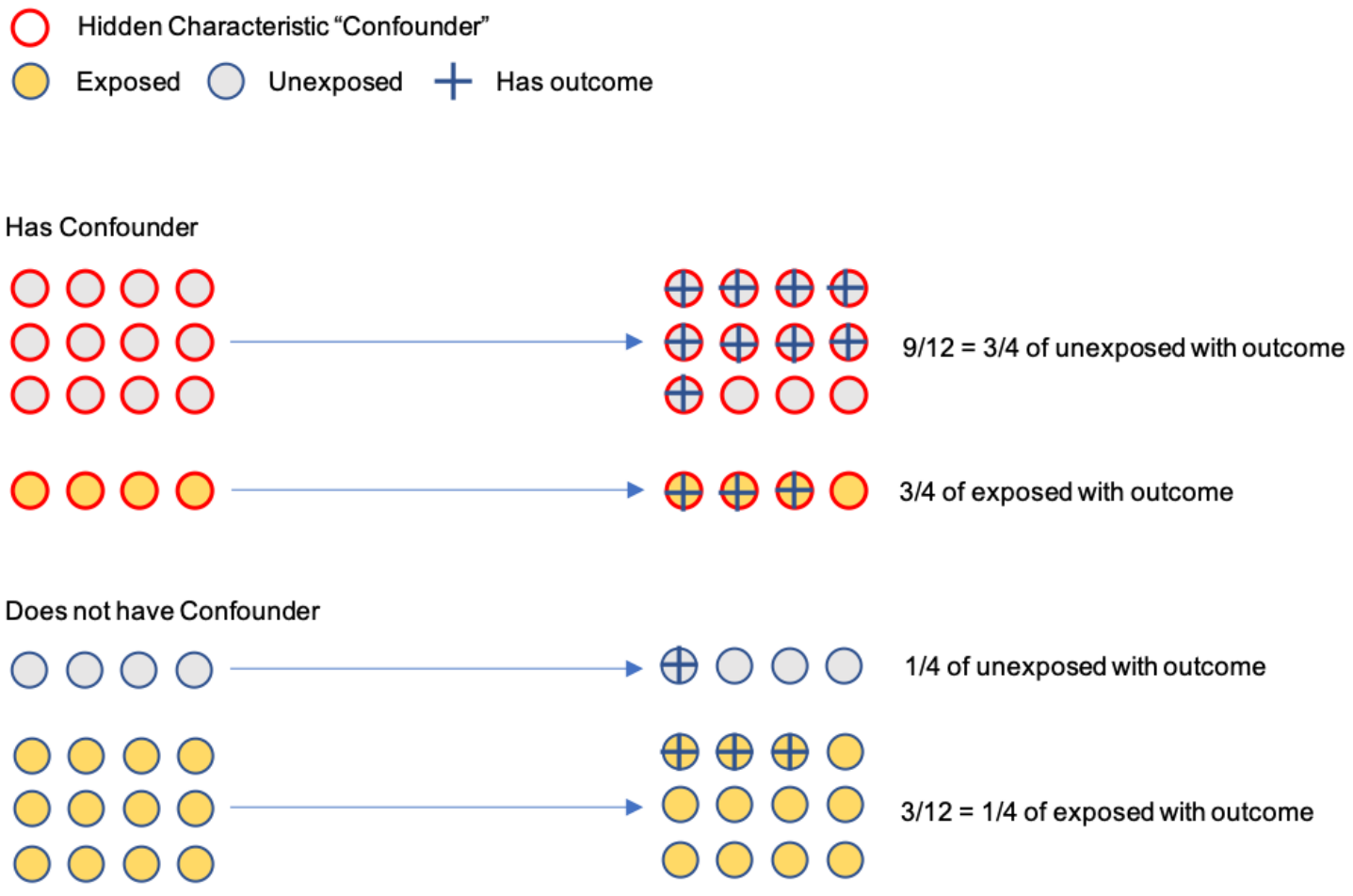
If you saw the result below, you would conclude that the exposure is protective: the proportion of participants with the outcome is much greater among the unexposed than the exposed.

Confounding occurs when there is a hidden (and hopefully measured!) factor, indicated by the red outline. The distribution of the hidden factor differs among the exposed and unexposed, and among those that have and don't have the outcome.

Biases: Confounding and Stratification (3/5)



Biases: Confounding and Stratification (4/5)



Biases: Confounding and Stratification (5/5)

If we have measured the confounder, then we can do what is called a **stratified analysis**: look within the strata of the confounder and assess the relationship between exposure and outcome separately.

You can see that within a given strata, the proportion who have the outcome is identical among those exposed and unexposed. There is no relationship between exposure and outcome.

Although we did this with a **binary** exposure and outcome, similar techniques exist for a continuous exposure and outcome.

Modeling: Intro

The basic motivation behind statistical modeling in epidemiology is that if you model the *data generating process* with some unknown parameters, then you can use observed data to estimate the unknown parameters of the data generating process.

$$Y = f(\theta)$$

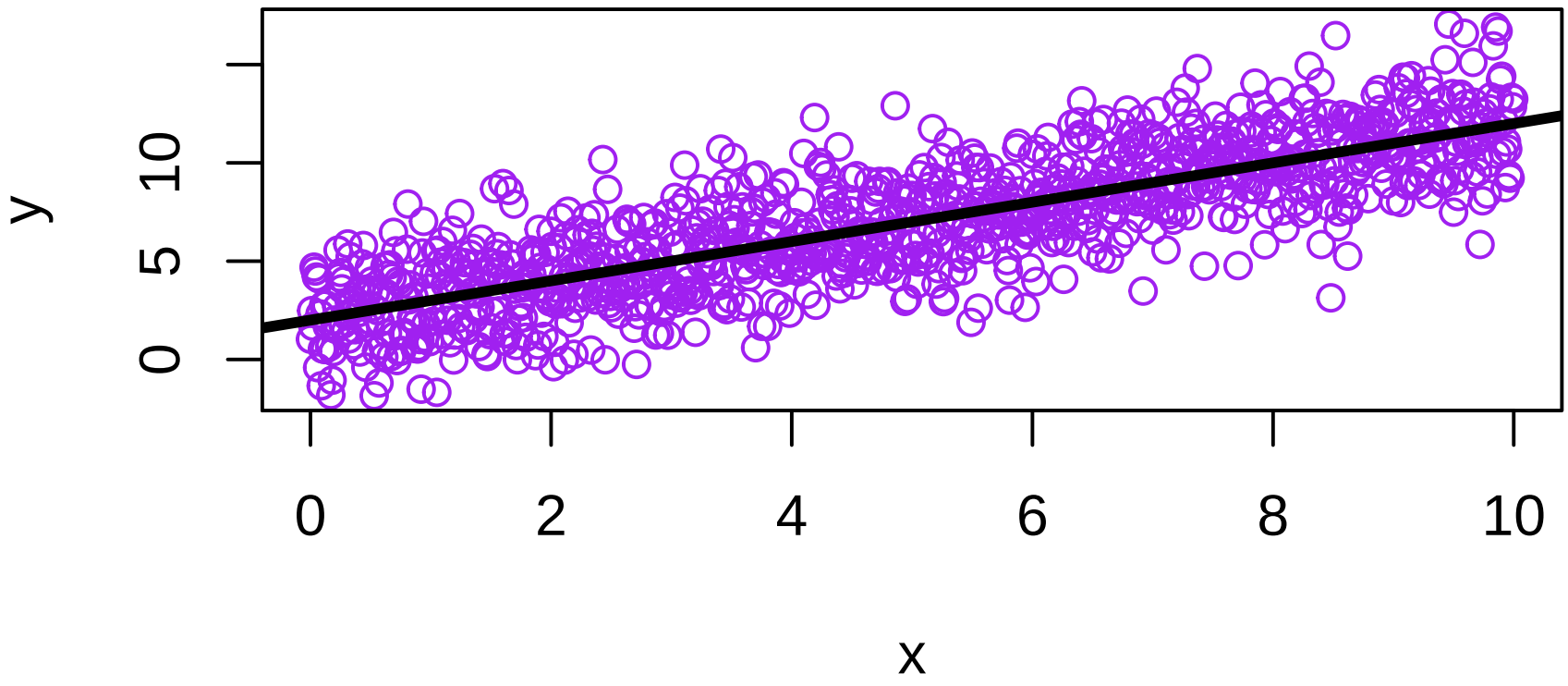
If we knew θ, then it might be interesting to generate some Y's. In fact, many mathematical modelers do this. But we are interested in the *inverse problem* – given Y, then what was θ?

The assumptions that you made in your study design go into $f(\theta)$ *– and in many ways* ***you*** *generated that data!*

Modeling: Linear Models

For most applications in epidemiology, the function f has some additional data that has been collected on individuals. Quite often, analyses use *linear models*.

$$y = \alpha + \beta x$$



Modeling: Randomized Controlled Trial

Consider a **randomized controlled trial**. We have an independent variable X that is 1 when the participant was randomized to treatment and 0 when they received placebo. We could ask,

- What is the mean of the outcome Y in the control group ($x = 0$)?
- What is the mean of the outcome Y in the treatment group ($x = 1$)?
- What is the *difference* between the mean outcome Y in treatment compared to control?

Modeling: Linear Models

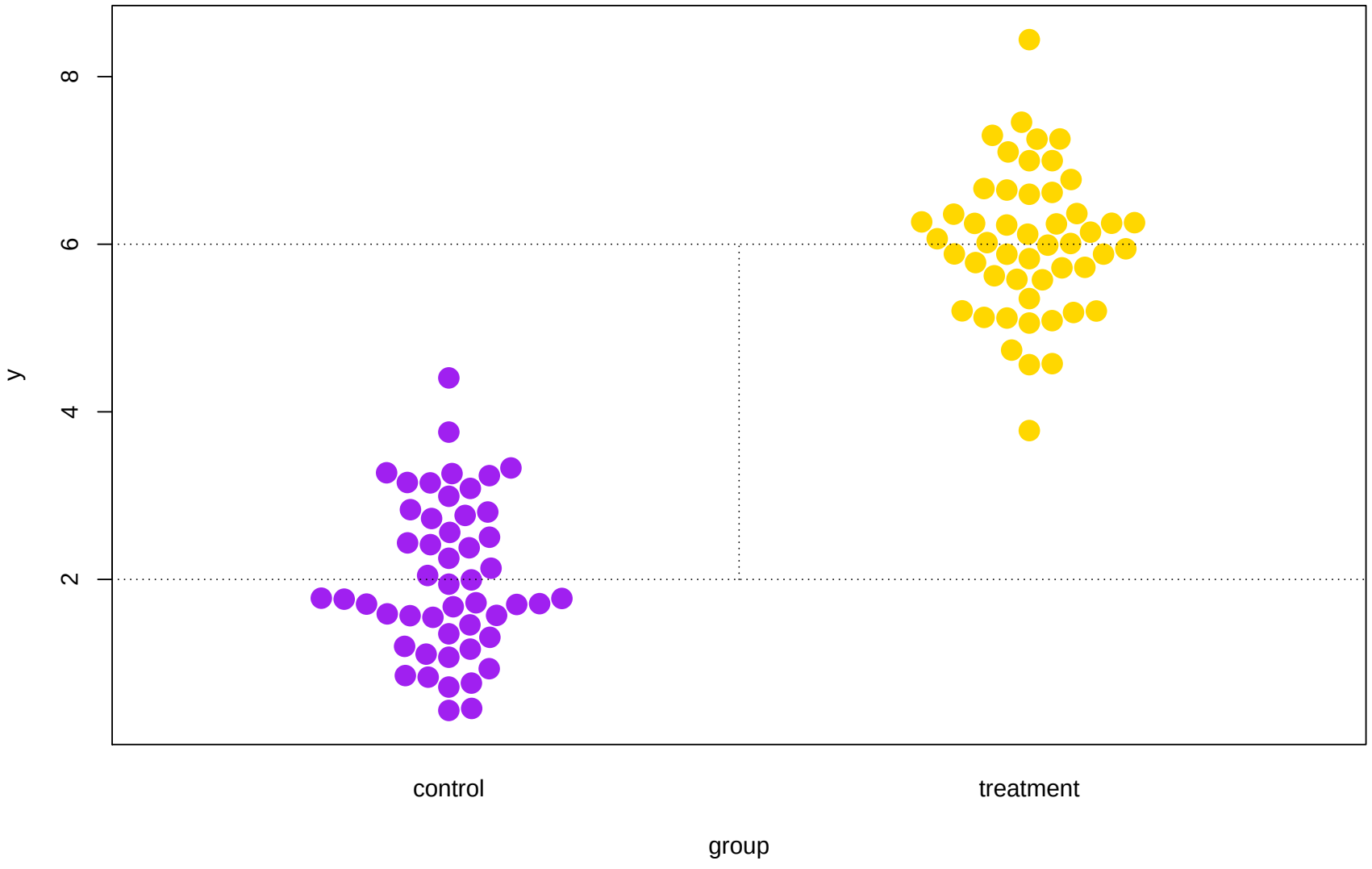
Thinking about both the **data generating process** and the **hypothesis that we want to answer**, we can parametrize a model.

$$Mean[Y|X = x] = \alpha + \beta x$$

- What is mean outcome Y among controls ($x = 0$)? $\rightarrow \alpha$
- What is mean outcome Y among treated ($x = 1$)? $\rightarrow \alpha + \beta$
- What is *difference* between the mean outcome Y in treatment compared to control? $\rightarrow \beta$

Often for this type of model, the error is assumed to be normally distributed. This means that the *average* outcome Y for someone with $X = x$, but there is some *random variation* around the mean that follows a normal distribution (bell curve).

Modeling: Visualization



Modeling: Example

Think back to our randomized controlled trial for women that have had a stroke, and let Y be whether or not they died, and X be their treatment group. We might start with the same linear model:

$$Mean[Y|X] = \alpha + \beta X$$

- What does α represent?
- What does β represent?
- Do you see any problems with how this model is parametrized?

Modeling: Example

Think back to our randomized controlled trial for women that have had a stroke, and let Y be whether or not they died, and X be their treatment group. We might start with the same linear model:

$$Mean[Y|X] = \alpha + \beta X$$

- What does α represent? **The proportion of participants that died in the control group.**
- What does β represent? **The difference in the proportion of participants that died in the treatment group.**
- Do you see any problems with how this model is parametrized? $Mean[Y|X]$ **can be negative or above 1 but proportions cannot!**

Modeling: Generalized Linear Models Example

The solution is to *rethink* the data generating process and add a *link function* g.

From our example, it makes most sense to think of Y as a flip of a loaded coin (Bernoulli distribution), and how loaded the coin is depends on the group X.

$$Mean[Y|X] = P[Y|X] = g(\alpha + \beta x)$$

$$g(x) = \frac{e^x}{1 + e^x}$$

This is called *logistic regression*. Now $\alpha + \beta x$ explicitly modify the probability of having the outcome Y, and it must stay between 0 and 1.

Modeling: Generalized Linear Model Example

Now, we also have to rethink our interpretation of α and β. It turns out that (and you can show with algebra):

- e^{α} is the odds of Y among controls
- $e^{\alpha+\beta}$ is the odds of Y among treated
- e^{β} is the ratio of odds of Y comparing treated to controls

where odds $= \frac{p}{1-p}$.

$$e^{\beta} = \frac{\text{odds of } Y \text{ in treatment}}{\text{odds of } Y \text{ in controls}}$$

Odds ratios are often reported in case control studies (because of the artificial distribution of cases and controls).

Modeling: Notes

- Other link functions and statistical distributions can be used to model all sorts of processes (e.g. count data).
- I've only presented examples where the independent variables (or "covariates") are binary 0/1. The same methods can be used when the independent variable is continuous.
- What I've presented so far is considered *parametric* modeling, and it is common in epidemiology. You may come across *semi-* or *non-parametric* modeling as you're reading literature. It is a way to make fewer assumptions about the data generating process, but will typically come with the tradeoff of increased variability in the result.

Inference

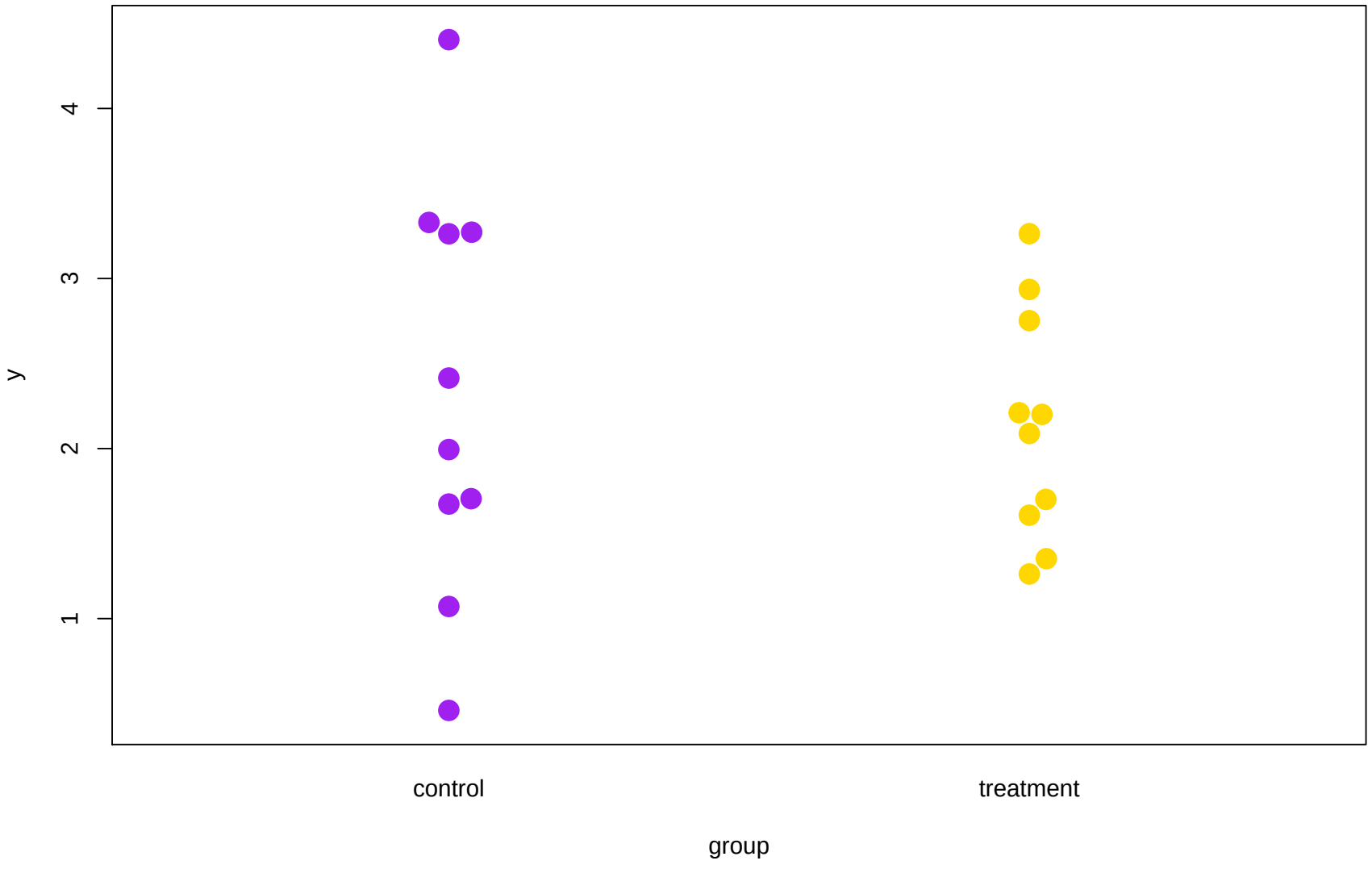
How do we actually estimate θ in $Y = f(\theta)$? There are a variety of techniques depending on what f is, but one thing is for certain – **there will always be uncertainty**.

We want to **infer** what θ is to some degree of confidence (typically 95% confidence).

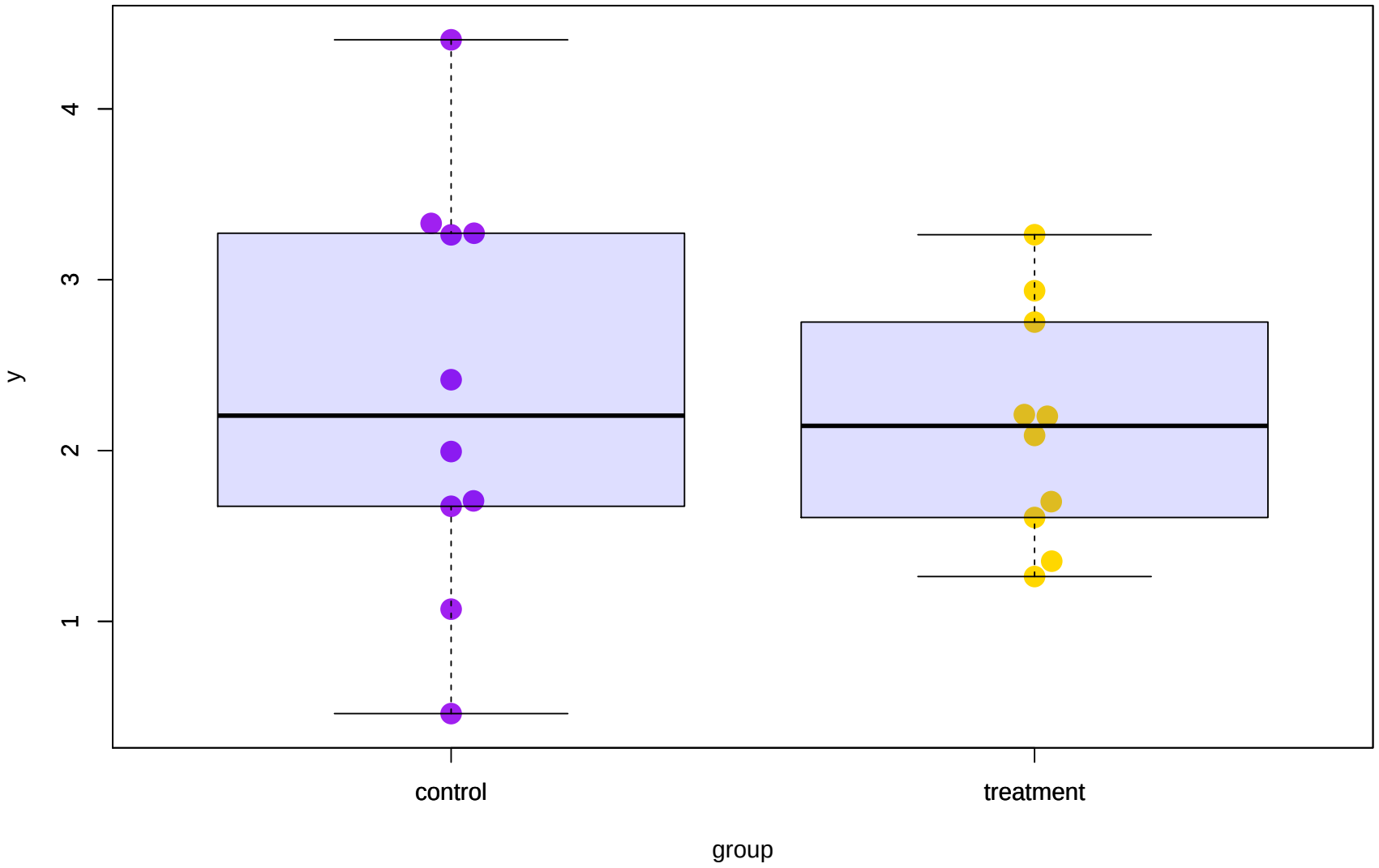
In medical literature, typically what you will see reported for a parameter are:

- Point estimate: the "best guess" for the parameter (in statistical terms, the "most likely" value)
- Confidence interval (95%): if this experiment were to be replicated 100 times, 95 of the confidence intervals constructed would cover the **true unknown parameter**
- p-value: quantifies the significance of an association

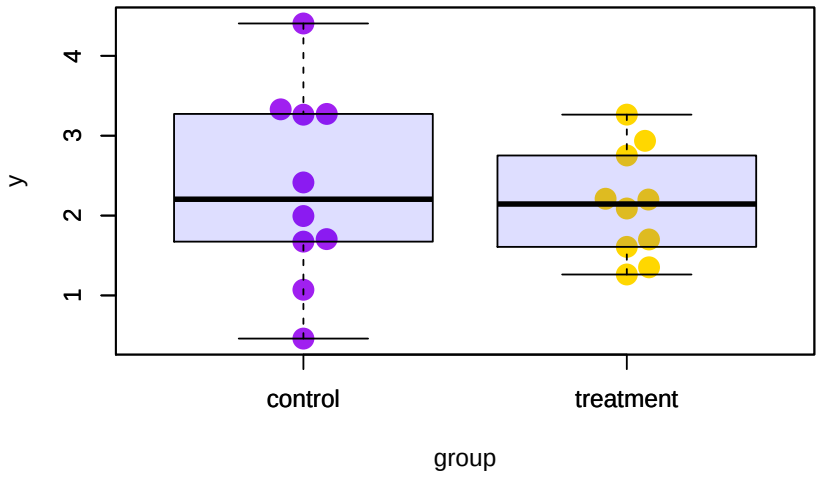
Inference: Simple Means

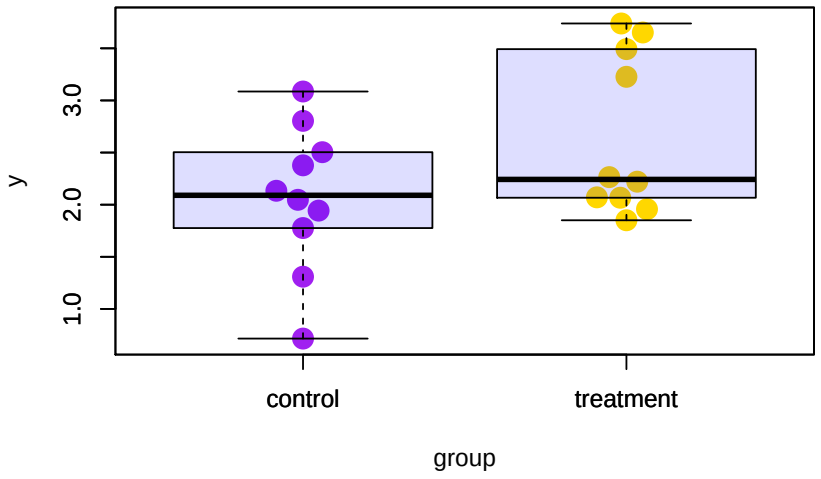


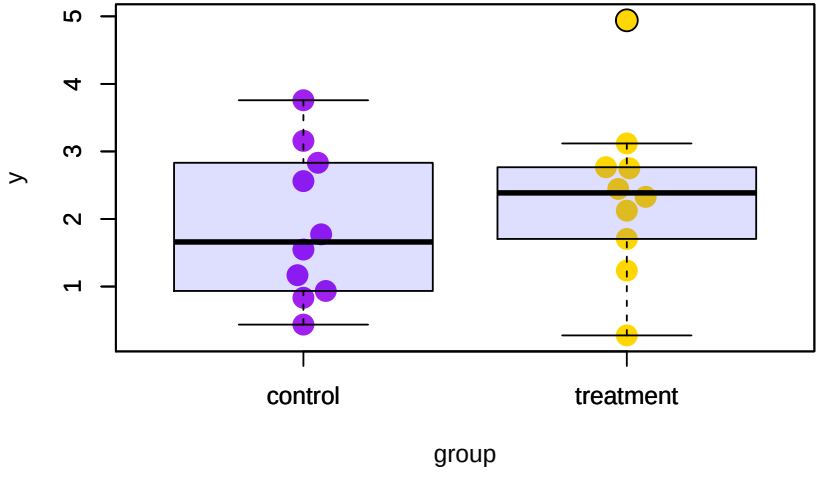
Inference: Simple Means

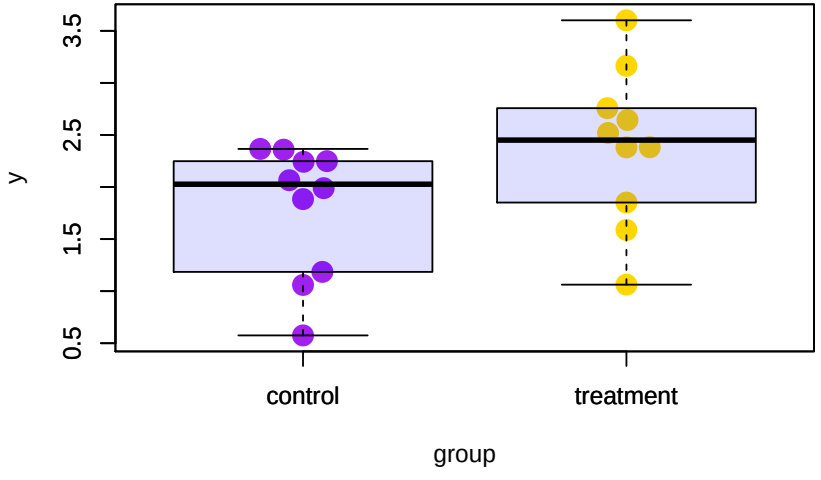


Inference: Small Effect Size, Small Sample Size

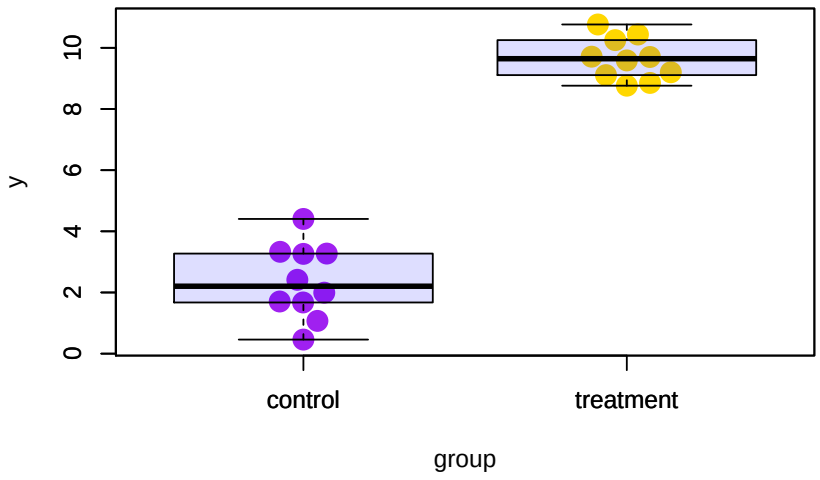


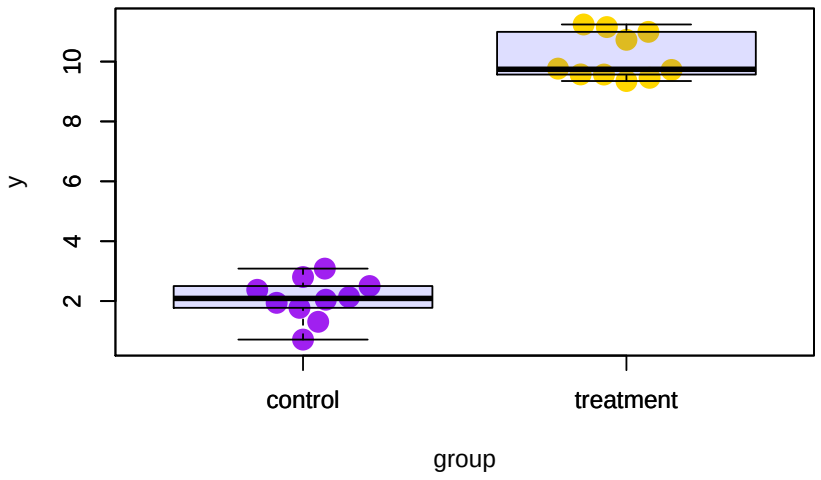


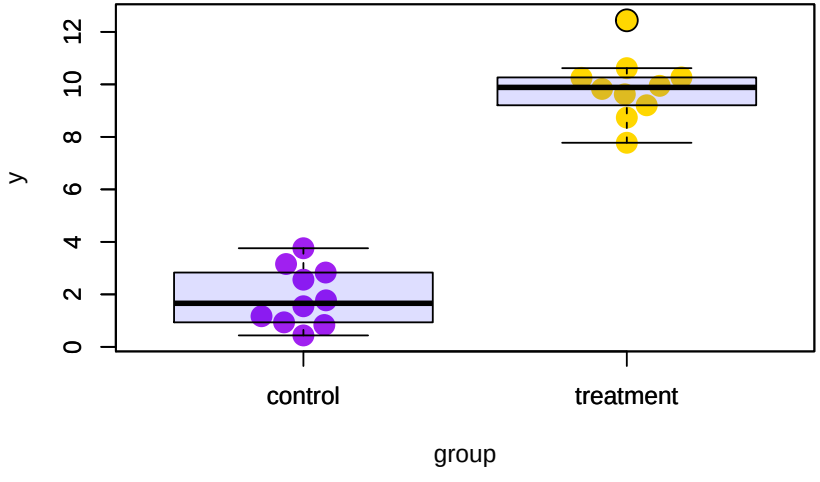


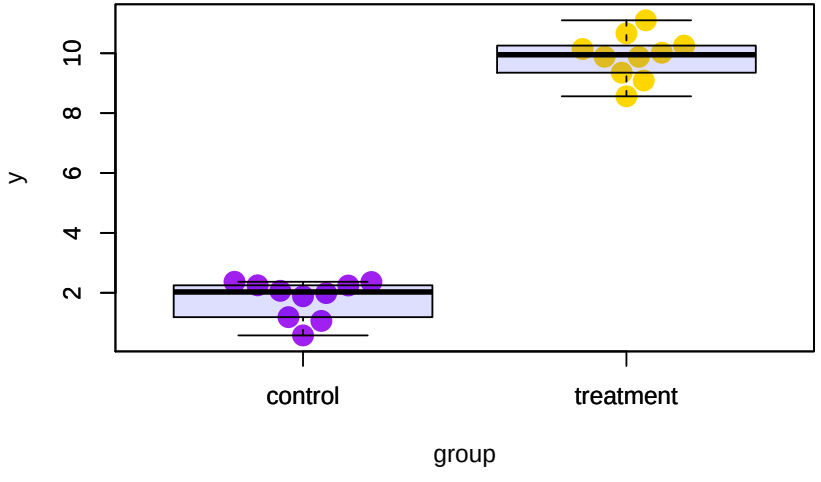


Inference: Large Effect Size, Small Sample Size









Inference: Small Effect Size, Large Sample Size