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)

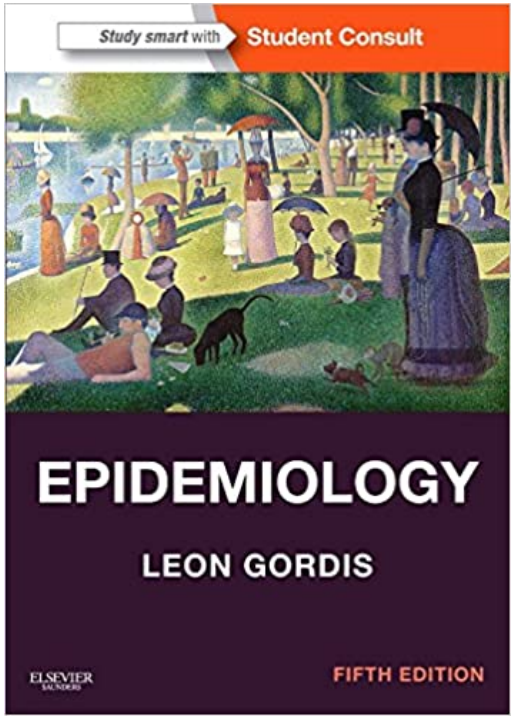
1. **Basics**. Identify key elements of an epidemiological study and how they relate to the scientific question
2. **Study Design**. Recognize the basic types of epidemiological study design and identify when each design is appropriate for the scientific question
3. **Bias**. Recognize sources of bias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

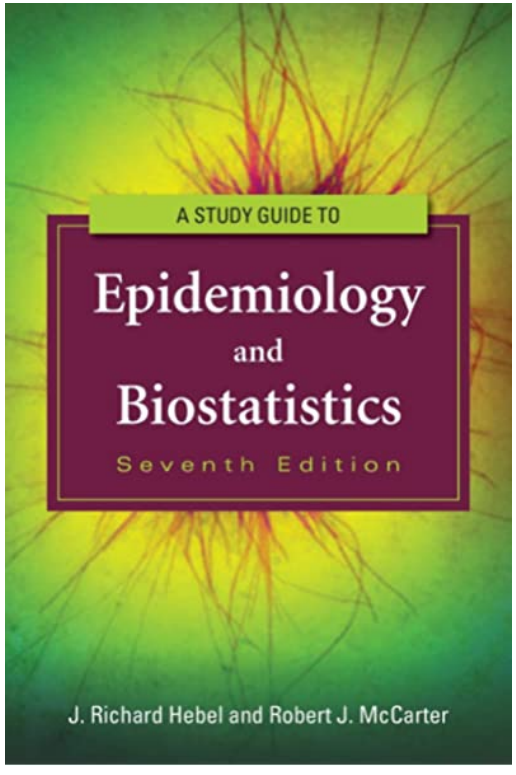
Learning Objectives (2/2)

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? Study populations should be representative of the target population that the study results will be generalized to.
- **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?

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
- **odds**: $\frac{p}{1-p}$ where p is a proportion

Think about denominators!

A common measurement in clinical epidemiology is **survival**, or time to some endpoint. We do not have time to go into details, but I recommend this source as a starting point.

Basics: Measures for Comparisons

Common measures of comparison between two exposure groups include functions of the measures we just discussed:

- **relative risk** or **risk difference**: comparing incidence of an outcome between two groups as a difference or a ratio
- **odds ratio**: comparing odds of an outcome between two groups as a ratio
- **hazard ratio**: compares the "hazard" of an outcome between two groups – used commonly in survival analysis

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 (1/2)

What are the exposure, outcome, and population for the following scientific question?

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

- Population:
- Exposure:
- Outcome:

Basics: Example (2/2)

What are the exposure, outcome, and population for the following scientific question?

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

- Population: **women age 45-60 who have had ischemic stroke**
- Exposure: **experimental treatment**
- Outcome: **death from ischemic stroke**

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:

Experimental Designs

- Randomized controlled trials

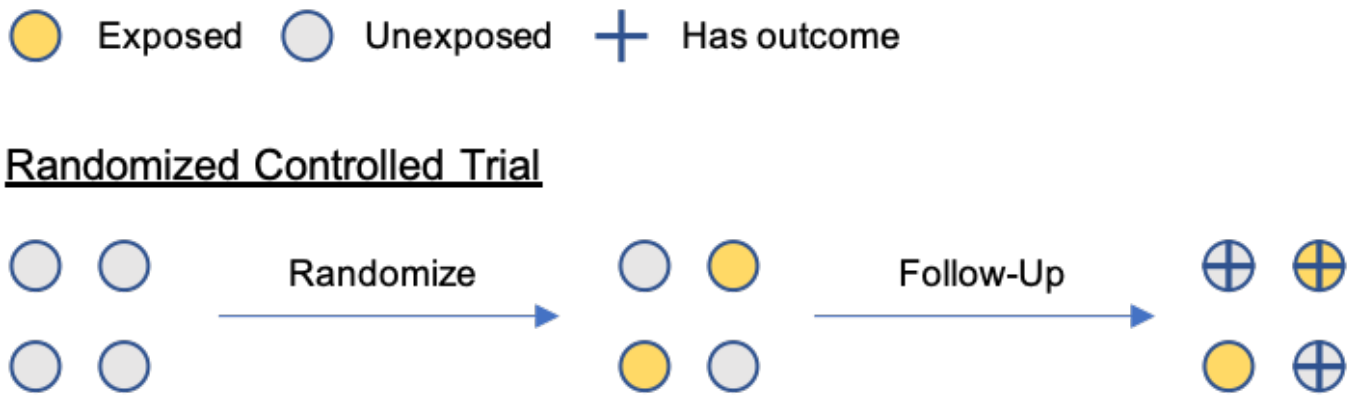
Observational Designs

- Cohort studies
- Case control studies
- Cross-sectional studies
- Case reports

Study Design: Randomized Controlled Trials

An experimental setting in which participants are enrolled and then *randomly* assigned to a treatment or a control and followed up over time to record outcomes.

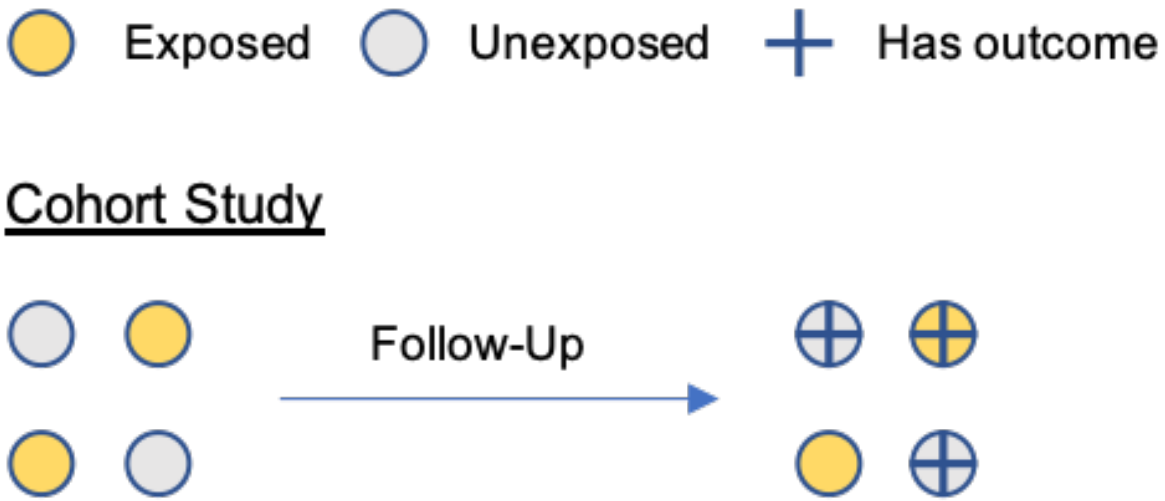
- Ideal for inferring causal relationships because there are no significant differences between treatment and control groups, both in terms of factors we can measure and factors we cannot measure (more later with confounding)
- They are expensive, time consuming, and not always feasible or ethical (think smoking)



Study Design: Cohort Study

An observational design where participants are selected based on their exposure status and followed up to record outcomes at designated time points.

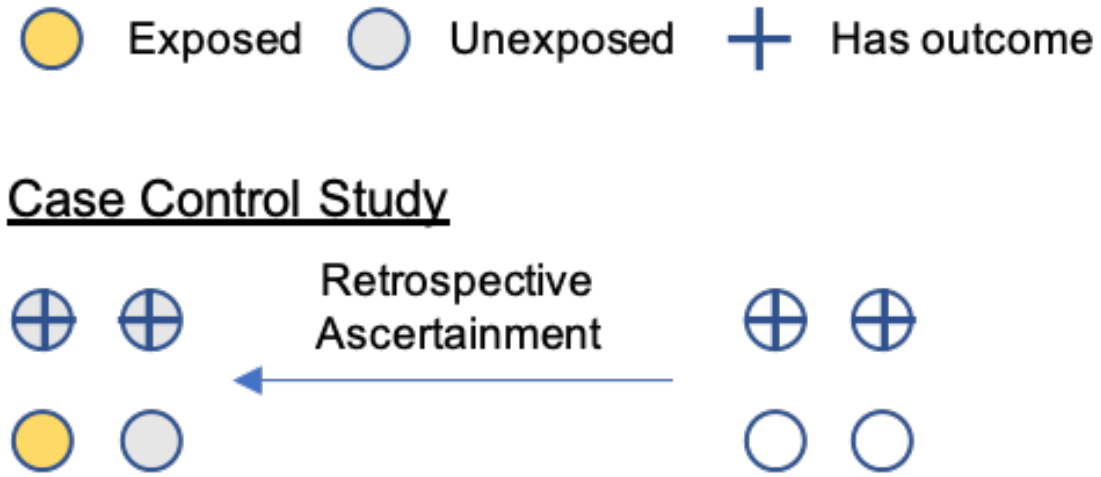
- Similar to a randomized controlled trial with one key difference: the exposures are not randomized! You may only **observe what is already happening**.



Study Design: Case-Control Study

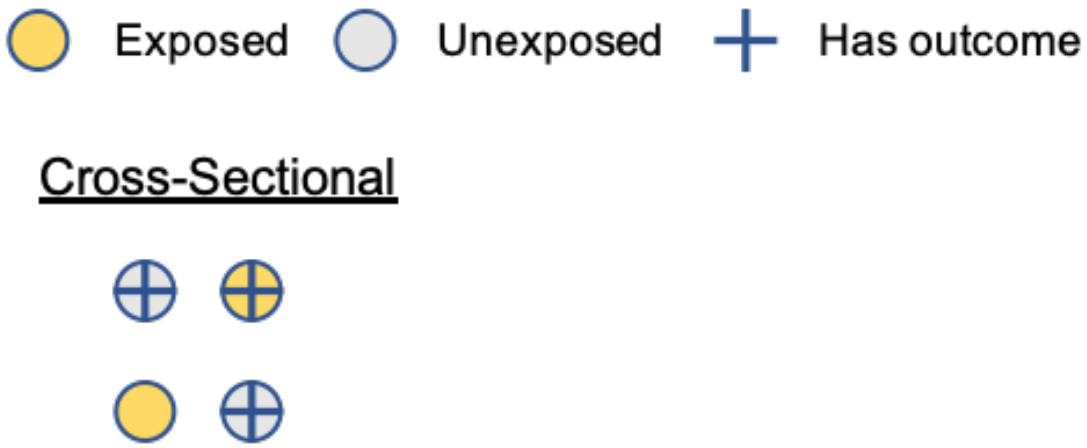
An observational design where participants are selected based on their outcome status and we inquire about exposure in the past. Think of it as the opposite of a cohort study, looking back in time.

- May be subject to biases like *recall bias* (more later)
- The artificial distribution of cases and controls means you cannot naively compute "prevalence" of the outcome or "relative risk"



Study Design: Cross-Sectional Study

An observational design where you measure exposure and outcome of participants at the same point in time (no temporal element).



Study Design: Case Report

An observational design where you report on the outcome status of one or a handful of interesting cases. "Anecdotal" evidence.

Study Design: Example

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

Which of the following designs would you prefer to answer this scientific question and why?

- 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

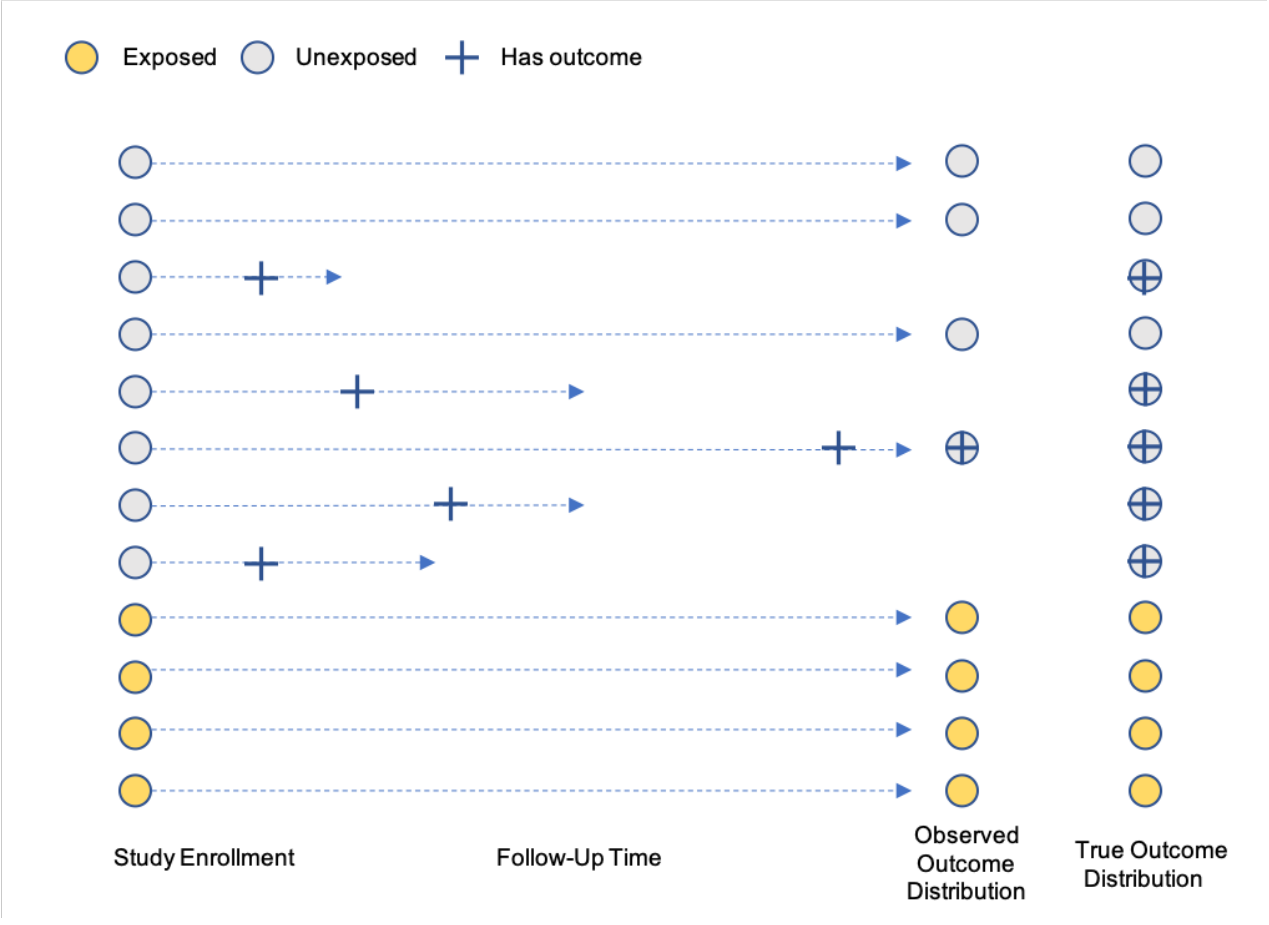
Examples of biases include:

- **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: 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?**

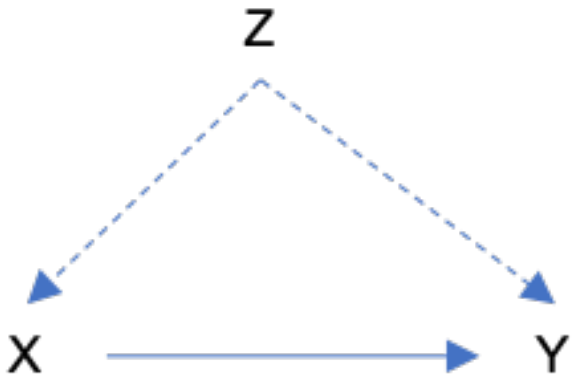
Consider the following sampling strategies:

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.

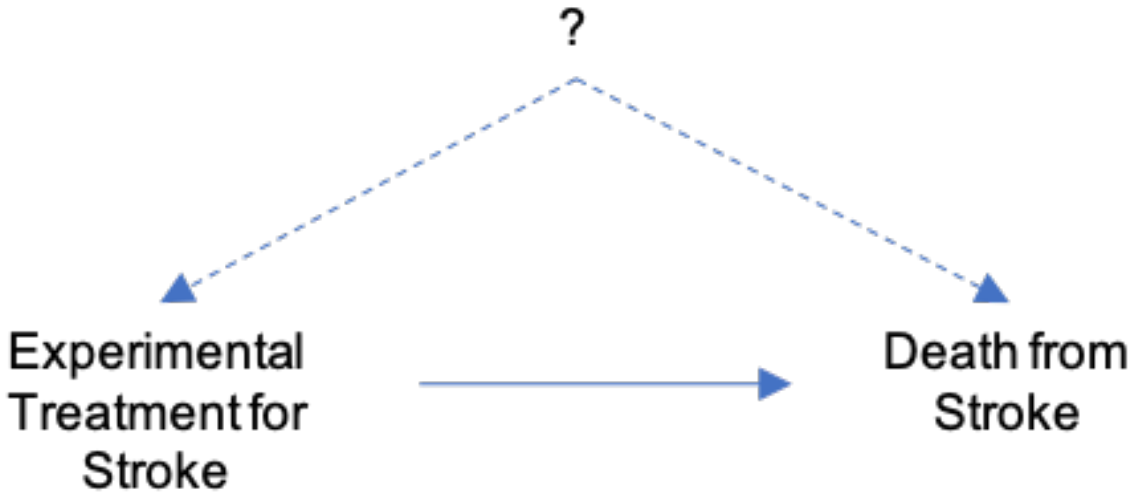


If we have **measured** Z, then there is hope that we can recover the true relationship between X and Y. If Z is unmeasured (which is often the case), it is much more difficult. In the case where Z is measured, there are standard techniques to "control" for Z.

Biases: Confounding Example

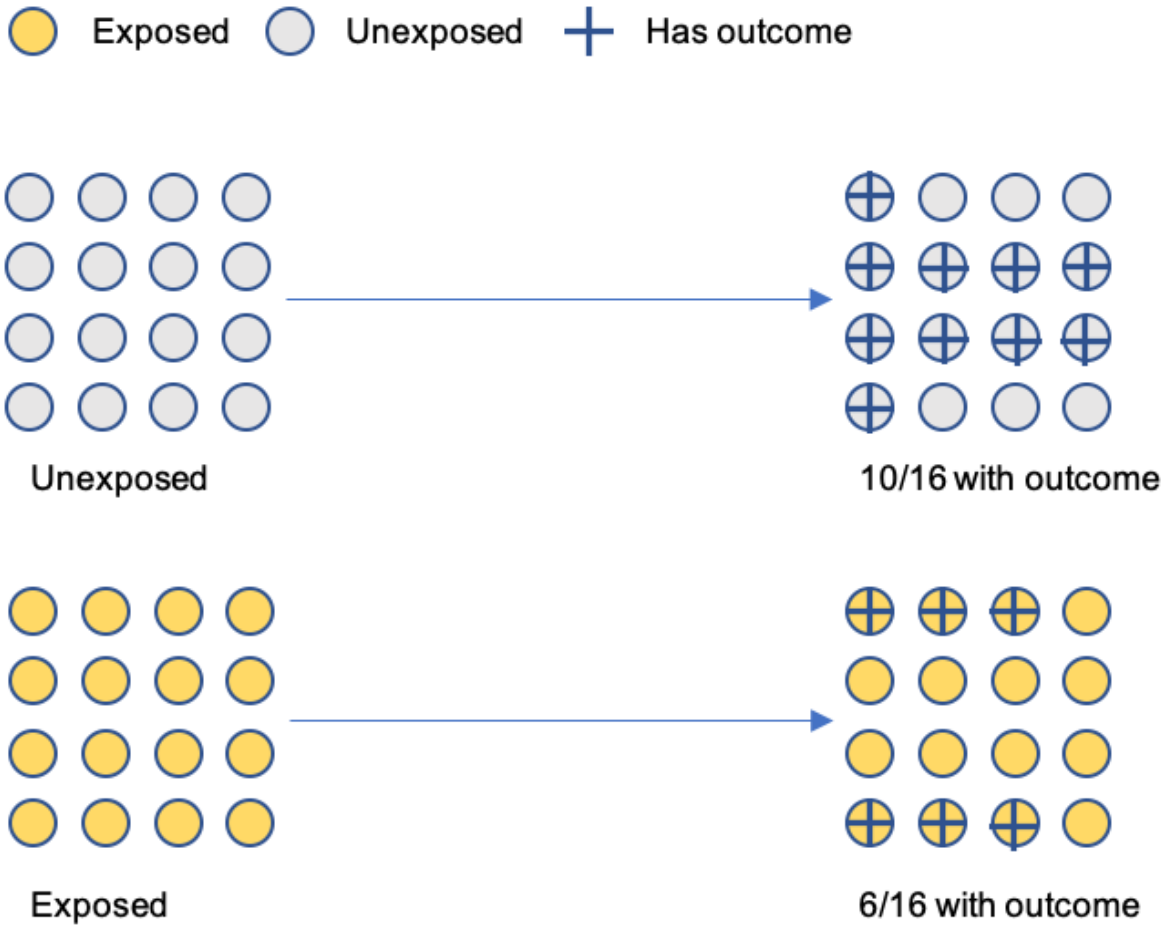
Again, recall our example inferential question: **What is the effect of an experimental treatment on mortality following ischemic stroke in women age 45 - 60?**

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.

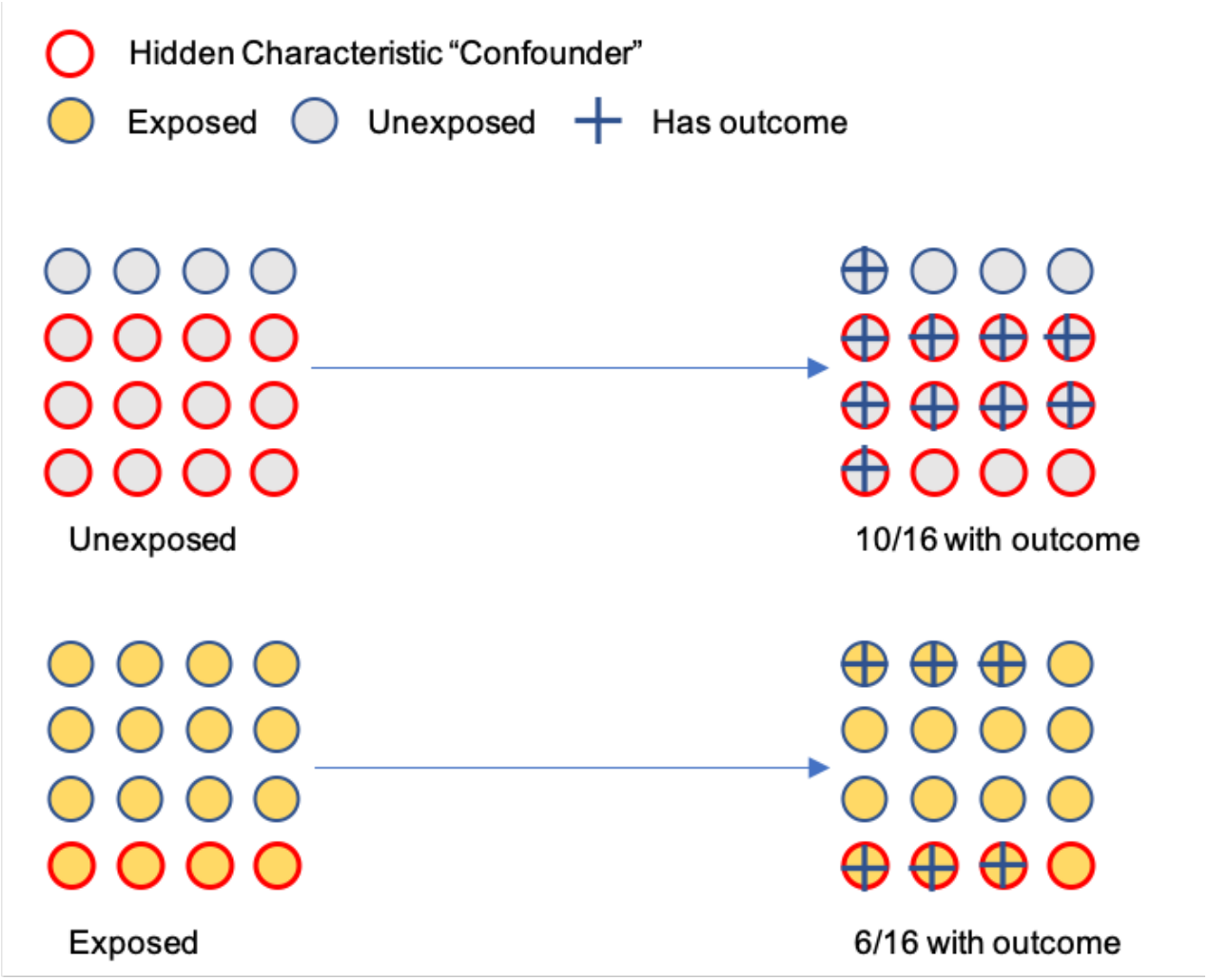


Biases: Confounding and Stratification (2/5)

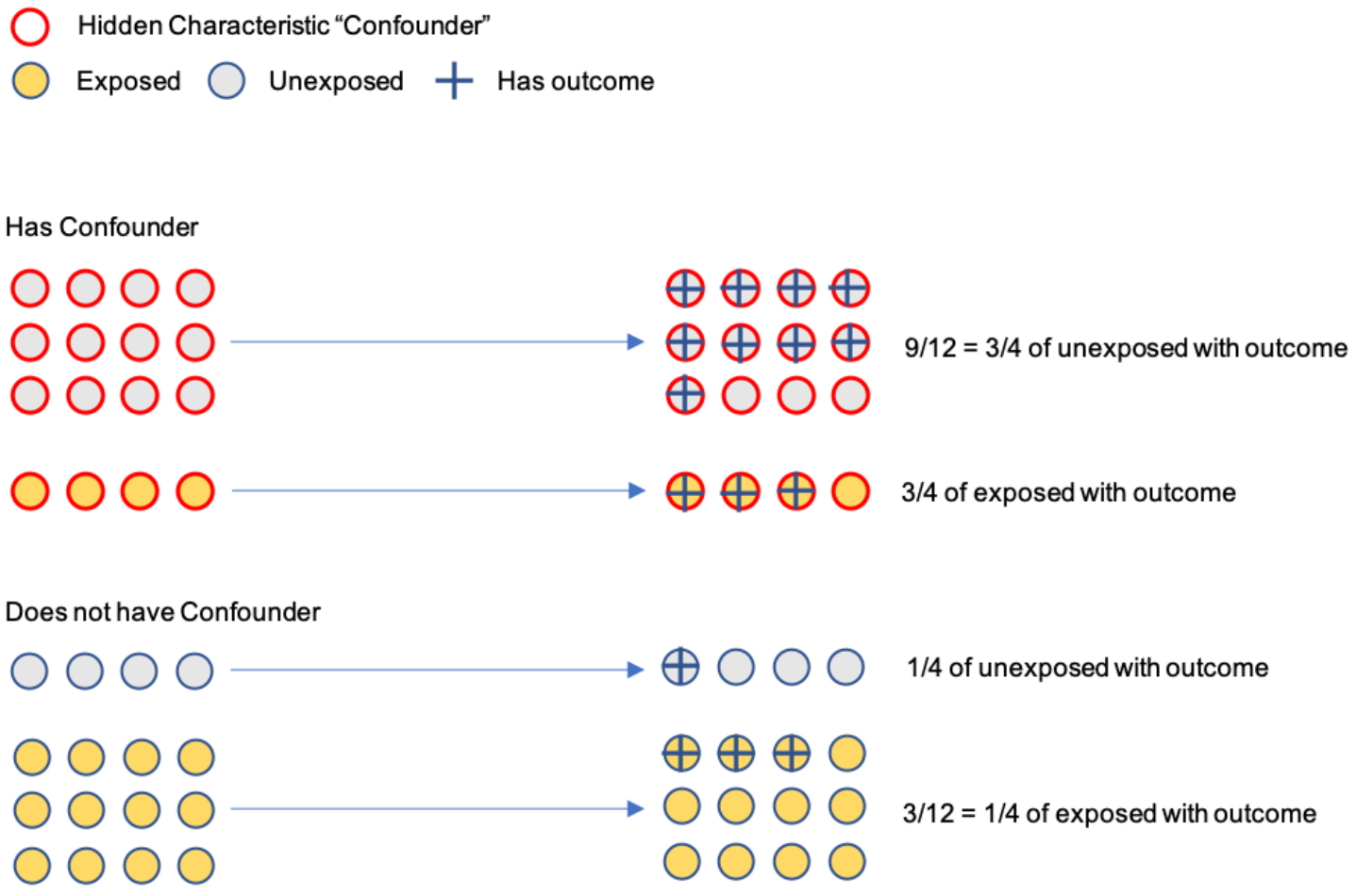
If you saw the result below, you would conclude that the exposure is protective: the proportion of participants with the outcome is much greater among the unexposed than the exposed.

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.

$$Y = f(\theta)$$

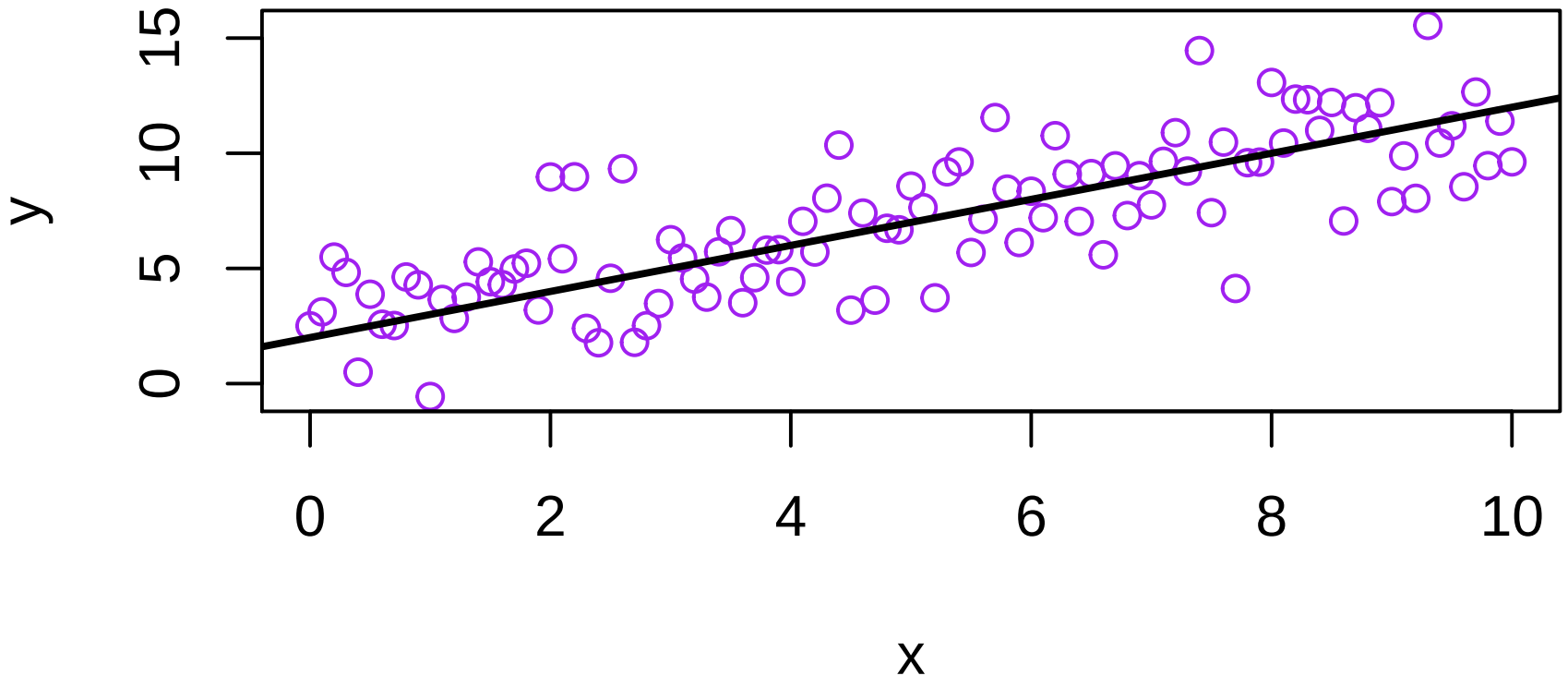
If we knew θ, then it might be interesting to generate some Y's. In fact, many mathematical modelers do this. But we are interested in the *inverse problem* – given Y, then what was θ?

The assumptions that you made in your study design go into $f(\theta)$!

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*, and find the "best fit" to the data.

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



Modeling: Generalized Linear Models

If you have data that looks like the previous plot with a continuous dependent variable, you're good to go with a basic linear regression.

Things get more complicated with different data types.

$$g(y) = \alpha + \beta x$$

Table 2: Common Mean Relationships and Distributions in Epidemiology

Data Type	Mean	Link Function	Statistical Distribution	Coefficients
continuous	linear	g(y) = y	Gaussian	risk difference
counts	log-linear	g(y) = log(y)	Poisson	log risk ratio
binary	logistic-linear	g(y) = logit(y)	Bernoulli	log odds ratio

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. The outcome Y is continuous. We could ask,

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

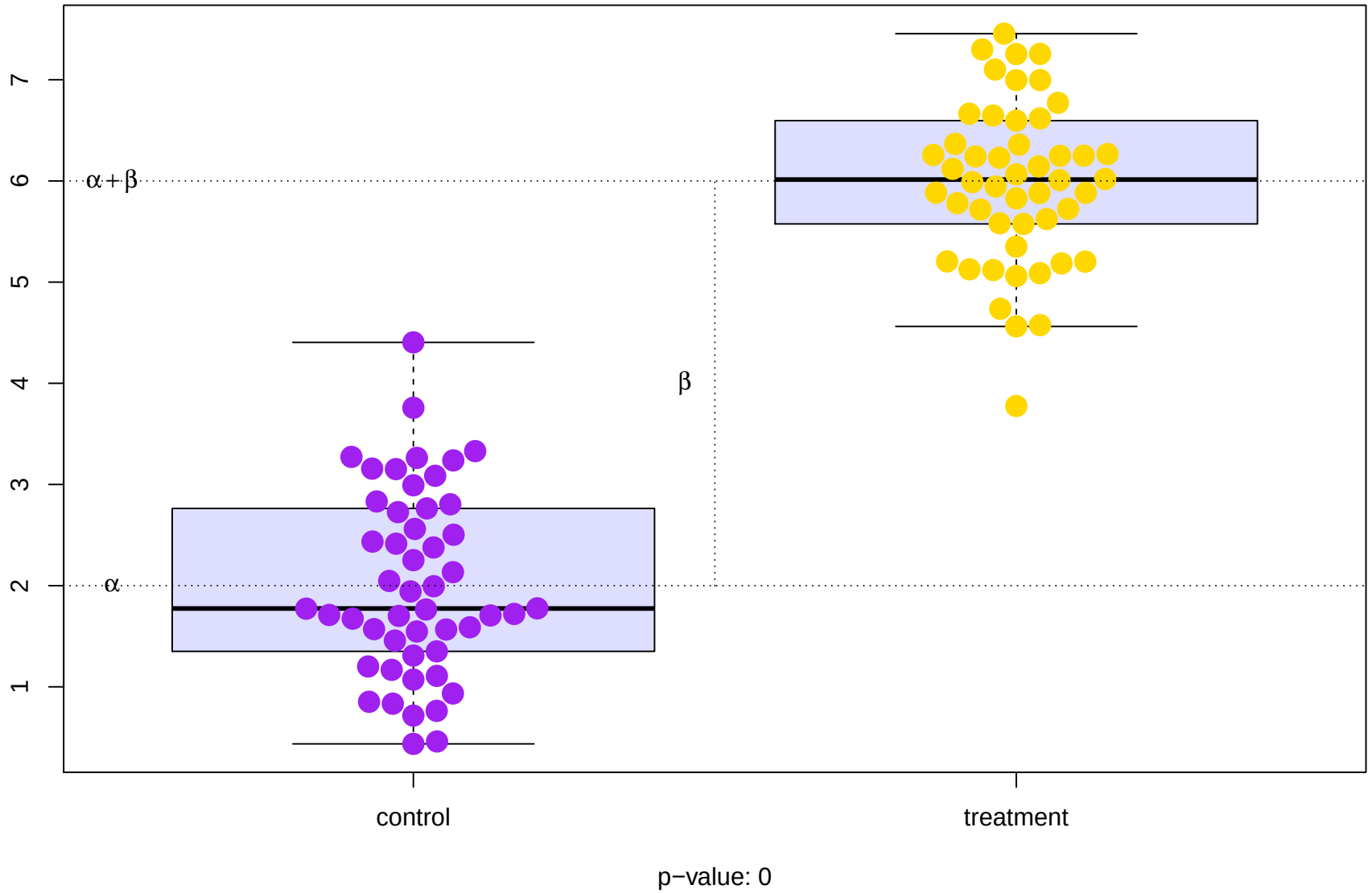
Modeling: Linear Models

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

Modeling: Visualization



Modeling: Example (1/3)

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:

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

Do you see any problems with how this model is parametrized?

Modeling: Example (2/3)

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:

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

Do you see any problems with how this model is parametrized? **Yes! The mean can go beyond the [0, 1] range, which doesn't make sense for a proportion.**

Modeling: Example (3/3)

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

$$\mathit{logit}(\mathit{Mean}[Y|X]) = \alpha + \beta X \rightarrow \mathit{Mean}[Y|X] = \frac{e^{\alpha+\beta X}}{1 + e^{\alpha+\beta X}}$$

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: Notes

- 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**.

Inference is about making our **best guess** about θ, and providing some degree of **confidence** in that guess (typically 95% confidence). This may also include a **p-value** for a hypothesis test.

In epidemiology, we only get to do a study once. **But in statistical inference, we use tools that are derived by *thinking about* what it would look like if we replicated the same experiment multiple times.**

Let's walk into these three elements in more detail for our stroke example

1. point estimates
2. confidence intervals
3. p-values

Inference: Point Estimates

In order to make the "best guess" for the e^{β} parameter, which represents the odds ratio of death after stroke comparing treatment to controls, we use information about the distribution that is specified (e.g. bernoulli distribution) and combine all of the observations together to figure out what value of β is the **most likely**.

This **most likely** value is reported as the **point estimate**.

"*We estimate an odds ratio of **0.6**, meaning that there was a 40% decreased risk of death in the treated group compared to untreated group.*"

Inference: Confidence Intervals

Confidence intervals (typically 95%) represent the degree of uncertainty that we have in our point estimate.

The precise statistical definition: if this experiment were to be replicated 100 times, 95 of the confidence intervals for the parameter constructed in these experiments would cover the **true parameter**.

The epidemiological interpretation: "the range of plausible values for the parameter".

"*We estimate an odds ratio of* ***0.6, with a range of plausible values being between 0.5 - 0.7****, meaning that there was a 40% (30% - 50%) decrease in the risk of death in the treated group compared to untreated group.*"

Inference: P-Values

If we have some estimate of θ, p-values make sense in the context of a **hypothesis test**. It might look something like this:

$$H_0 : \theta = 0 \quad H_1 : \theta \neq 0$$

This means: my **null** hypothesis is that $\theta = 0$. In linear regression, if θ was the slope of the line, that means *there is no slope* (or no association).

A **p-value** of 0.05 has the following interpretation: the chance of seeing a result this extreme or more extreme **if the null hypothesis were true** is 5%.

In our earlier example, you might write a statement like this:

"*We estimate an odds ratio of **0.6** with a p-value of 0.2. If there were truly no decrease in odds of death after stroke by taking the treatment, there is a 20% chance we would see an odds ratio at least this different from 0.*"

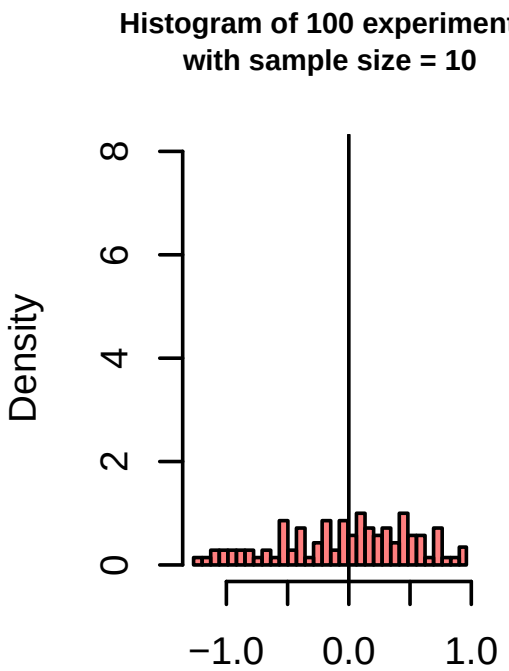
Sounds promising, right? ***Would you prescribe this drug your patients?***

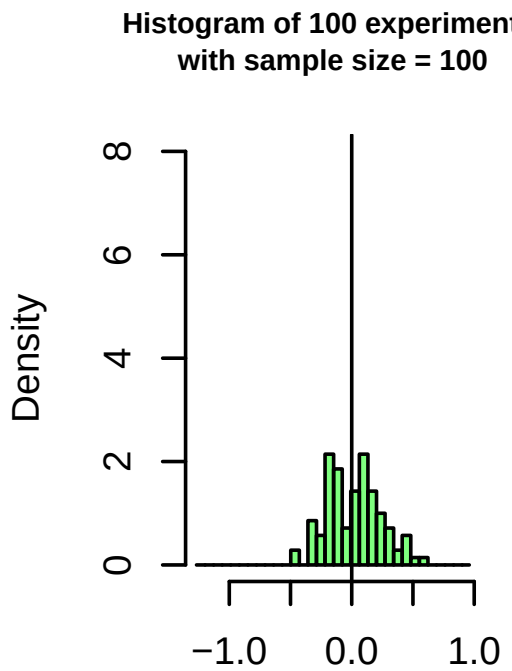
Inference: P-Values Example (1/3)

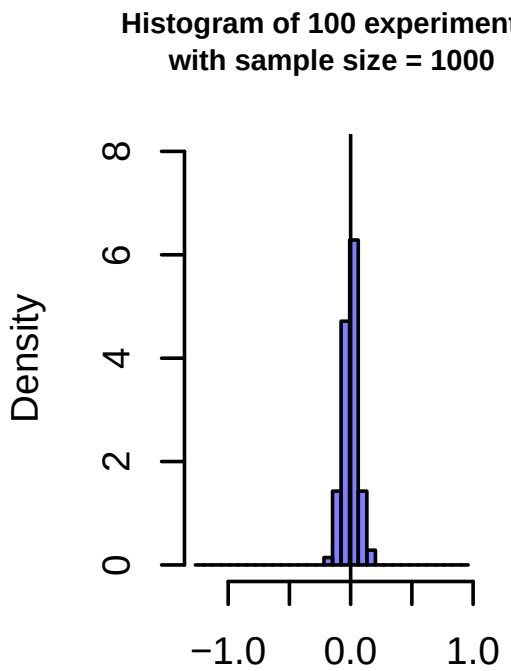
Let's explore p-values from another angle. Consider that the true decrease in odds of death is 0.0%, so the odds ratio is 1.0, or the log odds ratio is 0.

Consider running our stroke experiment 300 times, each time on a new set of people. In the first 100 experiments, we sample 10 people and randomize them to treatment or control. In the second 100 experiments, we sample 100 people, and in the third, 1000 people.

This is what the distributions of log odds ratios might look like.





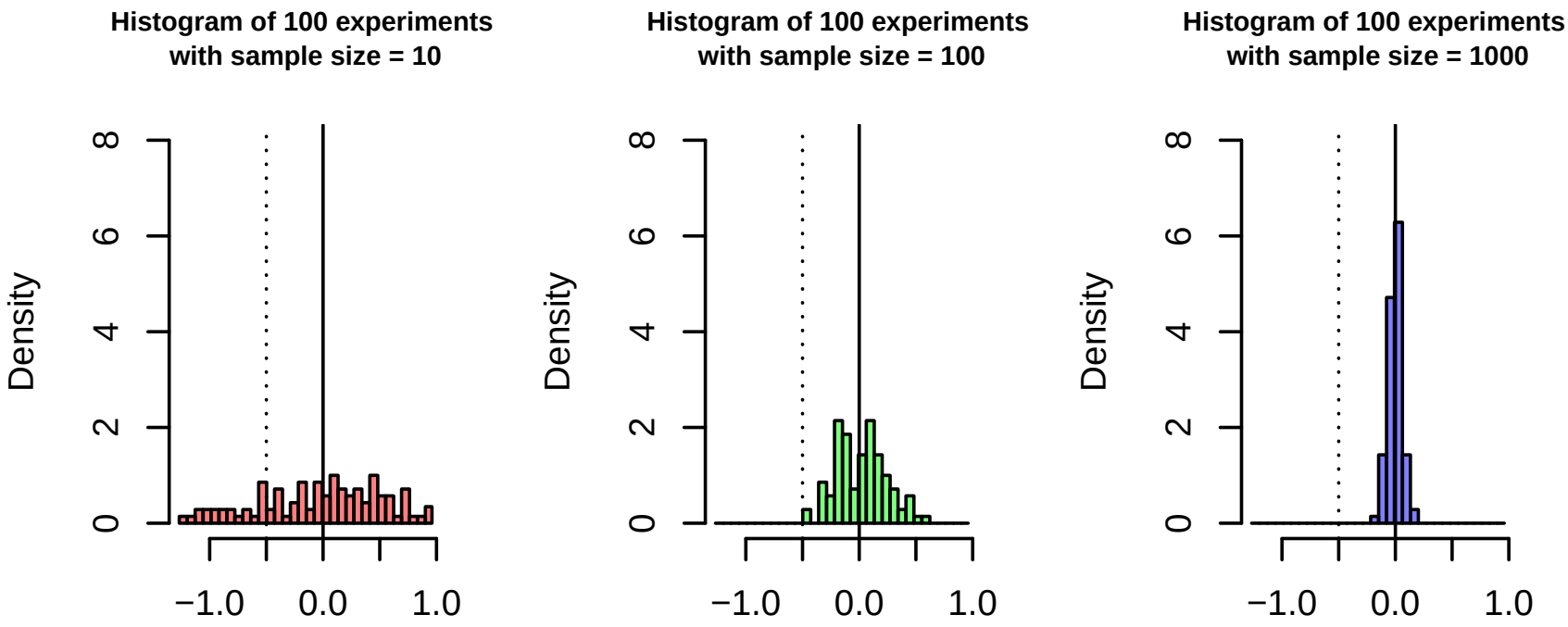


Inference: P-Values Example (2/3)

Now consider that you're back in the real world and you only get one shot to do a stroke experiment. You estimate a log odds ratio of -0.5 (odds ratio 0.6; 40% decrease in odds). If the experimental drug truly had no effect, how surprising would this estimate be?

The punch line: **it depends on the sample size of your experiment!** You get a smaller p-value with a larger sample size.

In this case, the effect is pretty large. A 40% decrease in odds of death!



Inference: P-Values Example (3/3)

This time when you do an experiment, you sample 10,000 people and you get an odds ratio of 0.99 with a p-value of 0.00001.

That is a *really small p-value*. If the experimental drug truly had no effect, this result would be very surprising. We're very sure this drug works. But do you really care about to prescribing a drug to your patients if it only translates to a 1% reduction in the odds of death?

P-values communicate statistical significance, which is important, but **point estimates and confidence intervals communicate clinical significance.**

Key Take-Aways

- A good study has meticulously defined its exposure(s), outcome(s), and population(s).
- Always be on the lookout for hidden biases and question whether or not the study authors accounted for them.
- It is important to consider the data generating process when parametrizing a statistical model.
- Point estimates (or "effect sizes") and confidence intervals are the most clinically relevant quantities to consider. **P-values cannot communicate clinical relevance.**

Discussion Articles

Next week we will discuss the following journal articles:

1. Anang et al. 2014. Predictors of dementia in Parkinson disease: A prospective cohort study. *Neurology* 83 (14). https://doi.org/10.1212/WNL.0000000000000842.
2. Caunca et al. 2019. Measures of obesity are associated with MR markers of brain aging: The Northern Manhattan Study. *Neurology*, 93:e791-e803. https://doi.org/10.1212/WNL.0000000000007966.

Disscussion Articles: Questions (1/2)

1. What is the scientific question of interest? Cearly define the exposure, outcome, and study population.
2. What study design did the authors use to gather data, and was it appropriate? Were there alternative study designs that could have been used?
3. Are there any sources of bias introduced from the study design including in who was included in the study population or how the exposure and/or outcome were ascertained that hinder the authors' ability to answer their scientific question?

Discussion Articles: Questions (2/2)

4. How did the authors formulate their hypothesis into a statistical model? What assumptions did the authors make when they chose that statistical model? Do you think that the assumptions they made limit their ability to answer their scientific question of interest?
5. How did the authors interpret the results of their analysis? Did they emphasize clinical significance or statistical significance?
6. How did the authors frame the limitations and implications of their results? Were they cognizant of their assumptions and potential biases?

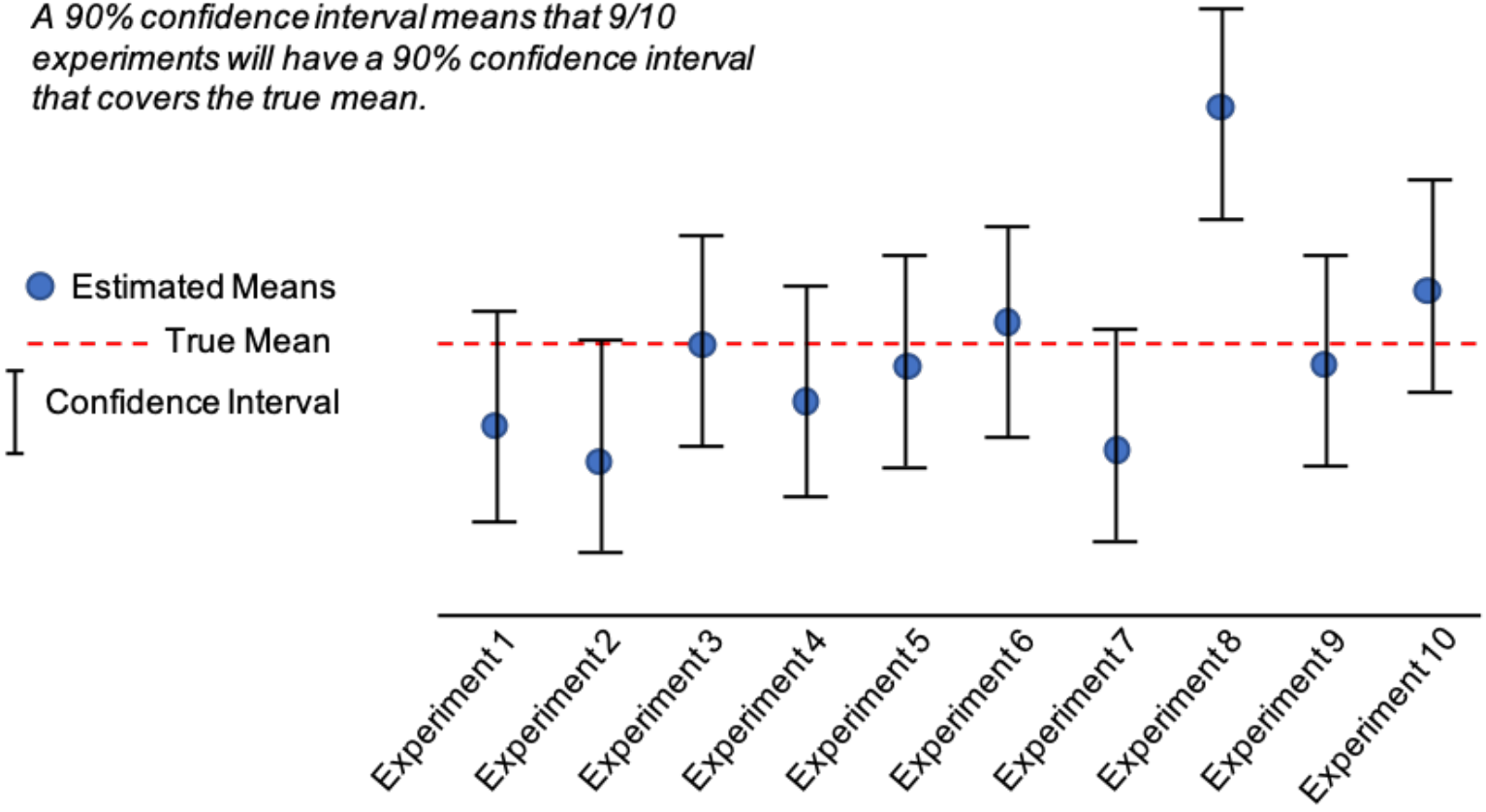
Thank you!

Questions?

Supplemental Material

Confidence Interval: Diagram

A 90% confidence interval means that 9/10 experiments will have a 90% confidence interval that covers the true mean.

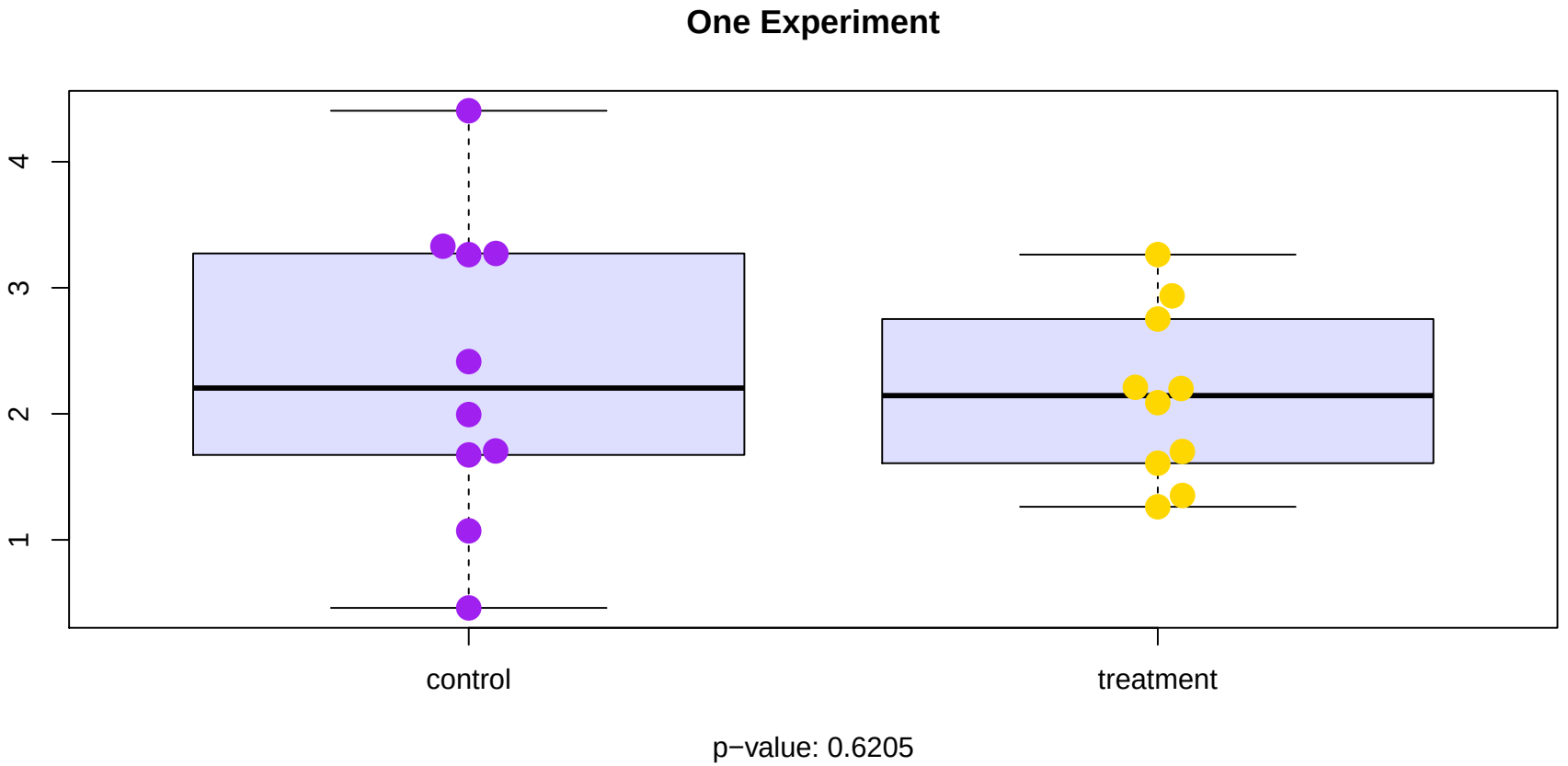


Inference: Experiments

The following slides show simulations for experiments with different effect sizes and sample sizes and how they impact the p-value for the test of the null hypothesis that the difference between the control and treatment group is 0.

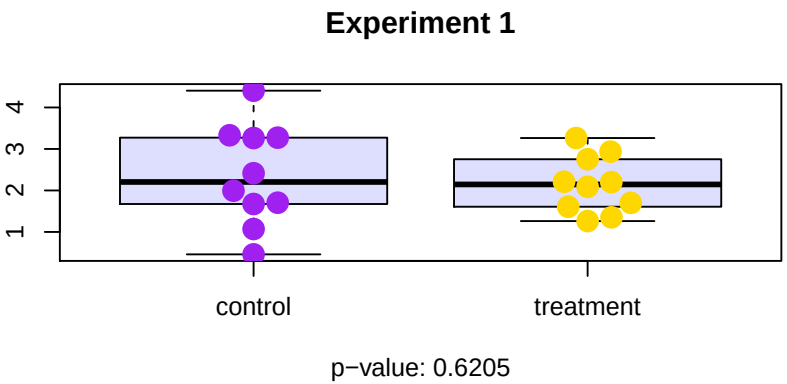
Inference: Simple Means

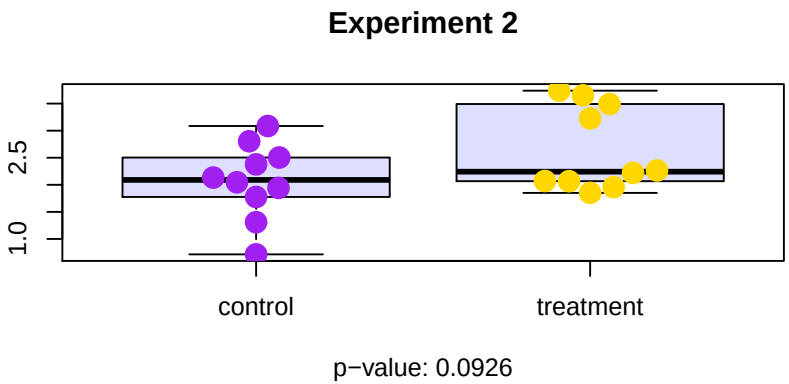
In this single experiment, the **true difference in means is 0.5**: treatment = 2.5, control = 2.0. The sample size is 10.

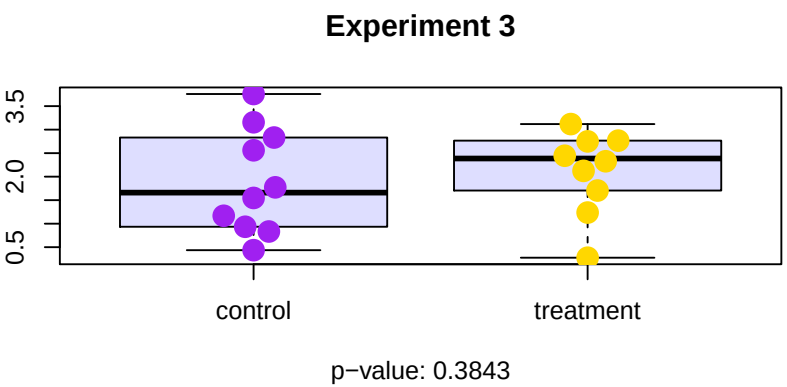


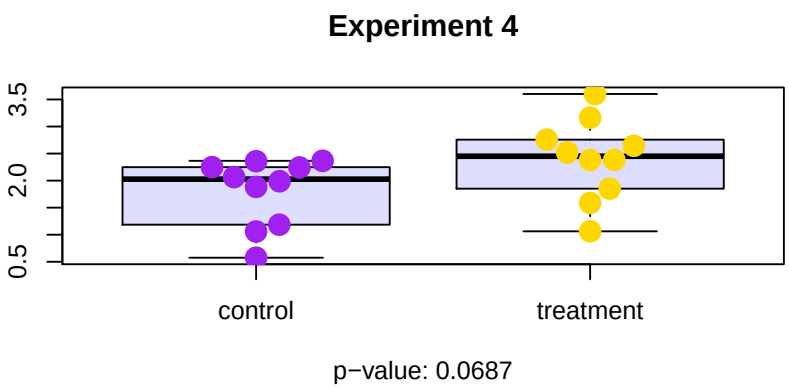
Inference: Small Effect Size, Small Sample Size

Here are four experiments instead of 1.



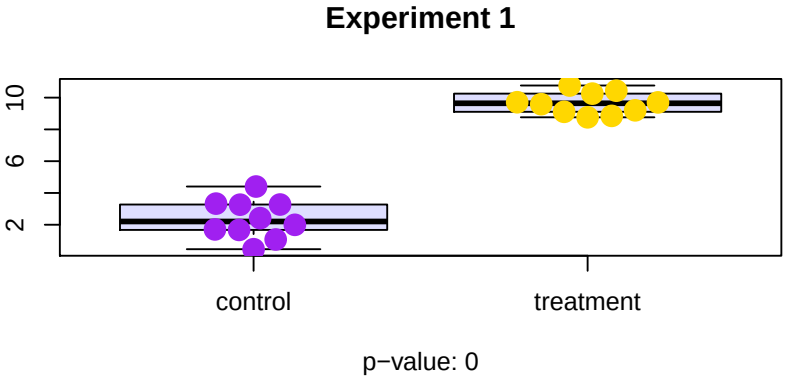


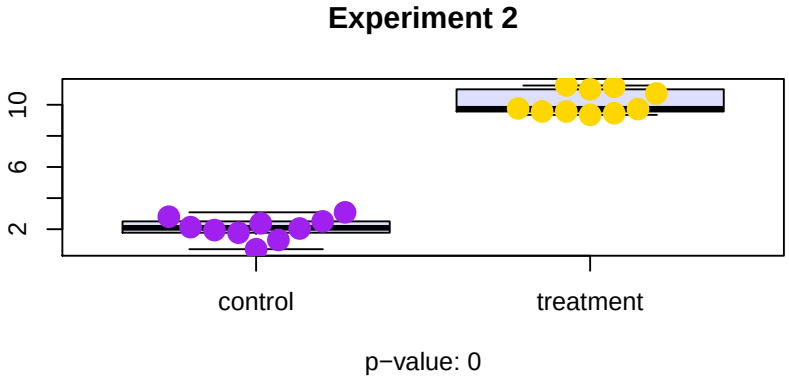


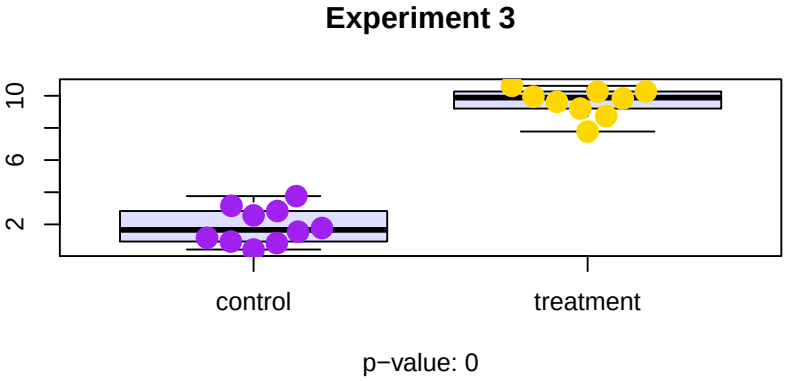


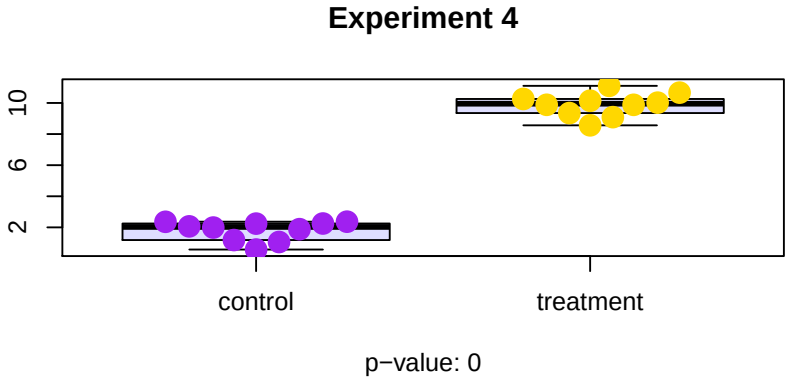
Inference: Large Effect Size, Small Sample Size

In these experiments, the **true difference in means is 8**: treatment = 10, control = 2. The sample size is 10. The large effect size translates to a small p-value, even with a small sample size.









Inference: Small Effect Size, Large Sample Size

In these experiments, the **true difference in means is 0.5**: treatment = 2.5, control = 2.0. The sample size is 10,000. The effect size is small, but the sample size is huge, leading to a very small p-value. **But is this clinically significant?**