

# Statistical Inference

Mitchell Hoffman  
UC Berkeley

# Introduction

Statistics: Making inferences about populations (infinitely large) using finitely large data.

Crucial for Addressing Causal Questions, e.g. :

- Does smoking cause cancer?
- Does pay for performance (P4P) lead to improved healthcare outcomes?

# Vaccinations and Primary School Graduation

- ▶ Are children who get vaccinations more likely to graduate primary than children who don't?
- ▶ Suppose you hear in the news: "The graduation rate for children with vaccinations was 75% compared to only 25% for children without."
  - What do we learn? 75% is much much larger than 25%.
    - What if the study had 4 kids in each group, with 3 of 4 vaccinated kids and 1 of 4 unvaccinated kids graduating from school?
    - What if the study had 4,000 kids in each group?

# Main Question in Statistics

- ▶ What can we learn from a given observed set of outcomes of a given sample size?
  - ▶ How confident can we be of inference based on 4 kids in each group?
  - ▶ How confident can we be of inference based on 4,000 kids in each group?
  - ▶ Confidence matters for evaluating whether a policy is working or not.
- 

# Review of Statistics Basics

- ▶ Suppose have data with N observations.
  - ▶ Mean (or average) =  $(1 / N) * (\text{Sum of all Data})$
  - ▶ Median: Middle value.
  - ▶ Mode: Most common value.
  - ▶ Minimum: Lowest value.
  - ▶ Maximum: Highest value.
- 

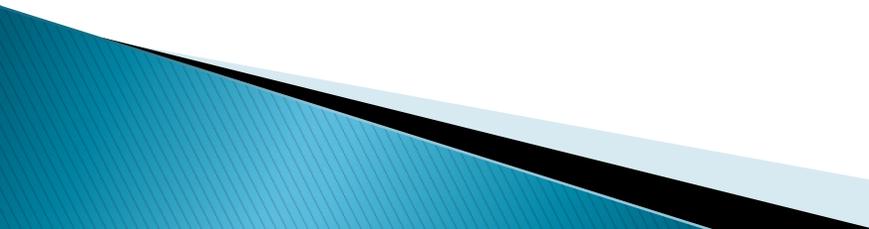
# Example

- ▶ Suppose we had data on the number of vaccinations provided to women from 5 different clinics: 23, 23, 30, 20, and 14.
- ▶ Mean =  $(1/5) * (23 + 23 + 30 + 20 + 14) = 22$
- ▶ Median = 23 (middle value)
- ▶ Mode = 23
- ▶ Minimum = 14
- ▶ Maximum = 30

# Hypothesis Testing

- ▶ In all hypothesis testing, we test an “alternative hypothesis” against a “null hypothesis.” In most real-world settings, the null hypothesis is one of no effect.

# Two-Sided Hypothesis Testing

- ▶ Let  $p$  be the probability that a mother attends all her recommended neonatal visits.
  - ▶ Null Hypothesis: Probability of attendance is the same for the Pay-For-Performance and the Control clinics.
  - ▶ Alternative Hypothesis: Probability of attendance is different between the Pay-For-Performance and the Control clinics.
- 

# One-Sided Hypothesis Testing

- ▶ Null Hypothesis : Probability of attendance is the same for the Pay-For-Performance and the Control clinics.
  - ▶ Alternative Hypothesis : Probability of attendance is higher for the Pay-For-Performance than for the Control clinics.
- 

# What to Use: One-Sided versus Two-Sided?

- ▶ Basic idea is that use one-sided only if have really strong idea about whether the effect will be positive or negative.
  - ▶ In practice, almost always use two-sided tests (this is what Stata gives as default). Two-sided tests make rejection harder.
- 

# Sampling Distribution and p-Values

- ▶ If your sample isn't the entire population, then your estimates will vary based on what sample you draw.
  - ▶ Suppose you want to know whether P4P clinics are providing more services than control clinics. The answer you will get to this question will depend on the number of clinics you sample.
- 

# p-values

- ▶ A p value = The probability that, assuming the null hypothesis were true, that one would get an estimate as or more extreme.
- ▶ Low p values means that it is less likely the null hypothesis is true.

# Statistical Significance

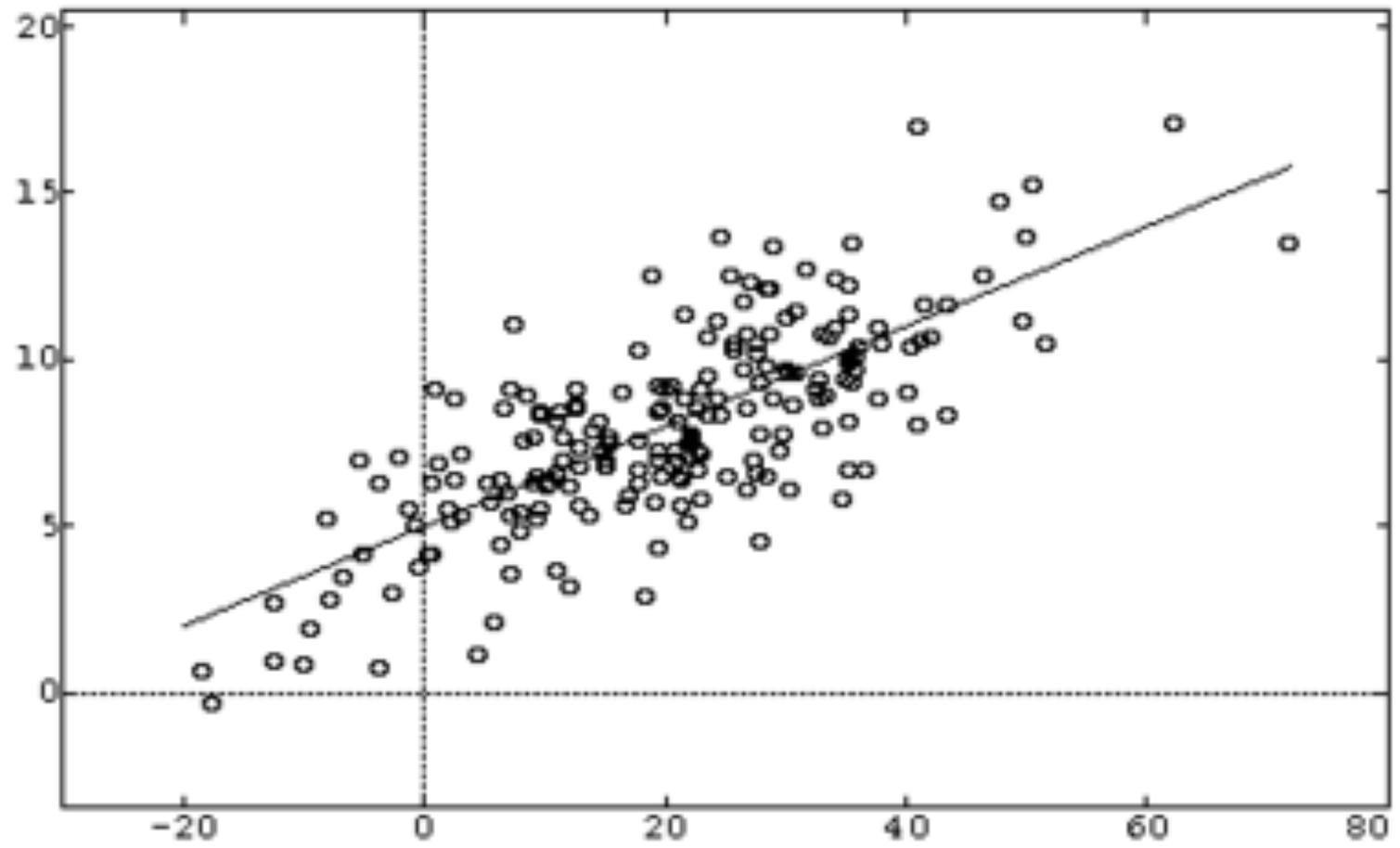
- ▶ A result is usually called statistically significant if it is unlikely to be due to chance alone.
  - By chance alone, it is possible that P4P clinics will do better than control. However,
- ▶ Most common standard for whether something is statistically significant is  $p < 0.05$ .

# Statistical Significance vs. Real-World Significance

- ▶ Just because you get statistical significance does not mean something is really significant for the real world. Mostly an issue when have lots and lots of data.
  - ▶ Suppose in a hypothetical example that P4P increased women receiving vaccines from 46 to 47 percent. With enough data, this will be statistically significant. However, this is a small change, and may be said to lack real-world significance.
- 

# Regression

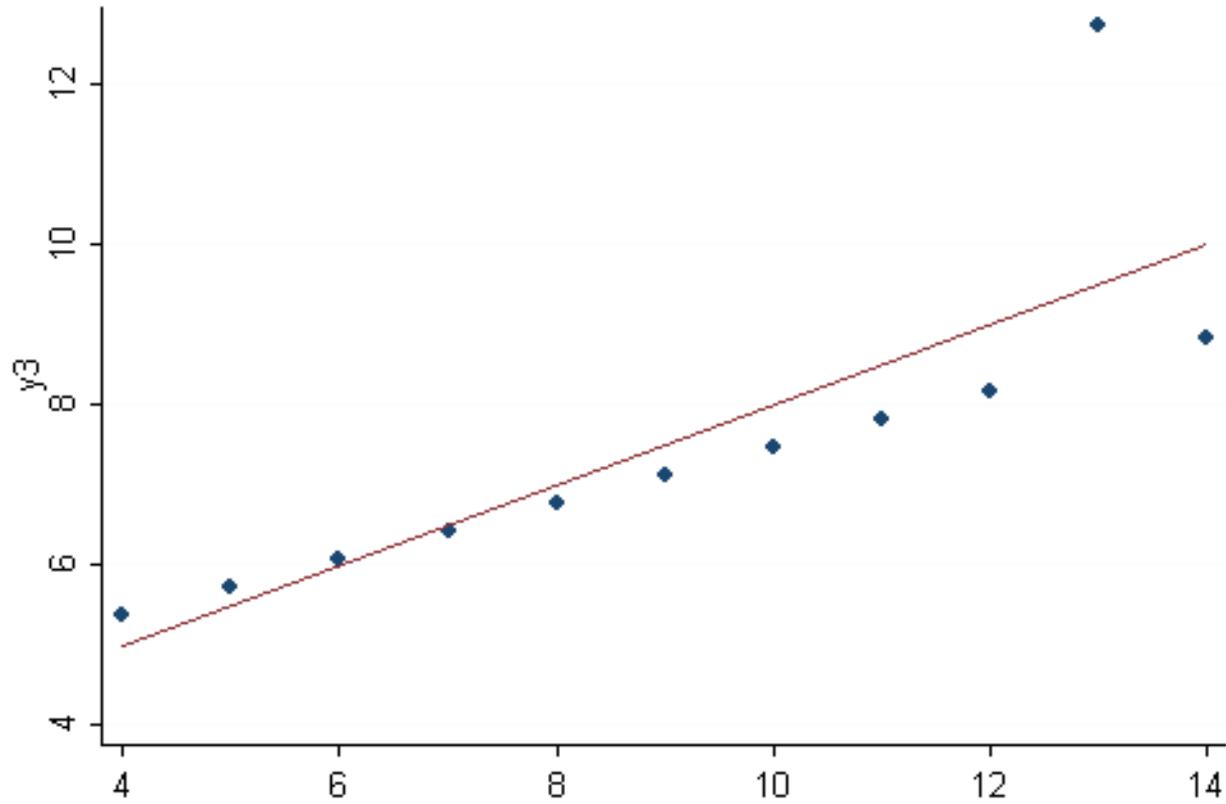
- ▶ Regression attempts to find the line that best fits the data. It offers the simplest and most common way to do hypothesis testing.



# Outliers

- ▶ A small number of “outliers” or unusual points can have a large effect on a regression.
  - Regression tries to find the line that best fits the data. One point way off the line can really shift things.
- ▶ Good idea to make a picture and do regression excluding the outliers.

## An Outlier



# Regression Equation

- ▶ Frequently, you will see equations like  
$$y = a + b \cdot x + u$$

Examples for pay for performance:

$y$  = Share of Mothers Receiving All Prenatal Visits.

$x$  = Whether the person was in treatment or control.

# Special Case of Regression: What if $y$ is a 0 or 1 variable?

- ▶ Suppose  $y$  is a 0 or 1 variable. Instead of share of women who get all prenatal vaccinations, suppose we wanted to model whether each woman got all the vaccinations.
- ▶ For this, it is best to use a logit or probit regression instead.
  - > Very easy to implement in Stata.

# Hypothesis Testing with Regression

▶  $y = a + b \cdot x + u$

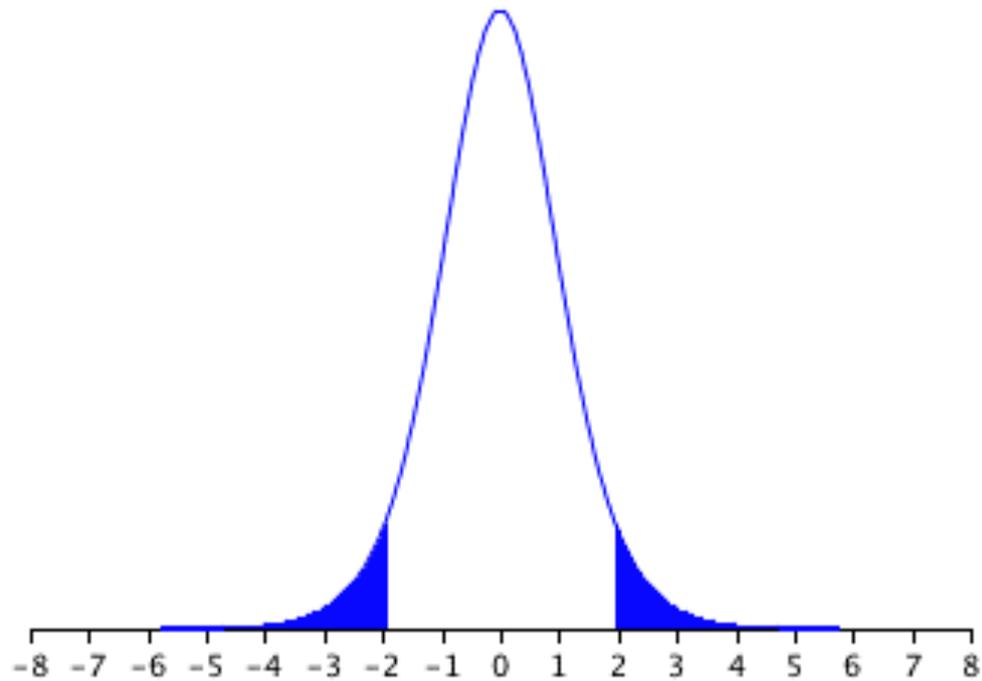
Null Hypothesis: P4P has no effect on share of women getting vaccinations (b equals zero).

Alternative Hypothesis: P4P affects the share of women getting vaccinations (b doesn't equal zero).

# t-statistics

- ▶ A t-statistic represents the sampling distribution for a regression under the null hypothesis.
- ▶ We use t-statistics to get from regression coefficients to p-values.
  - Stata does this for us.

t distribution with df = 8



df:

t:

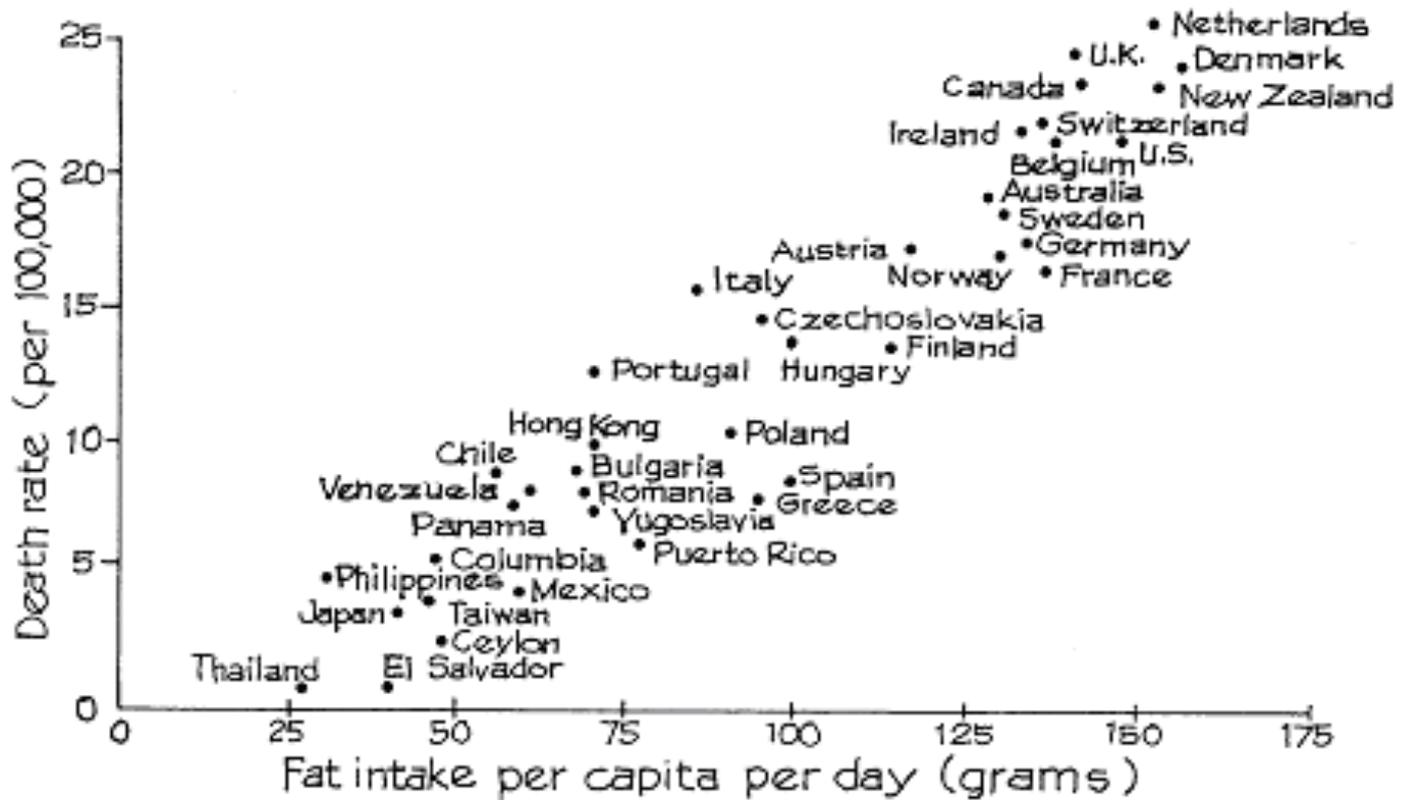
One-tailed test  Two-tailed test

Shaded area: 0.0857

# Correlation vs. Causation

- ▶ Even if a result is highly statistically significant, not necessarily the case that you have demonstrated true causation.

Figure 8. Cancer rates plotted against fat in the diet, for a sample of countries.

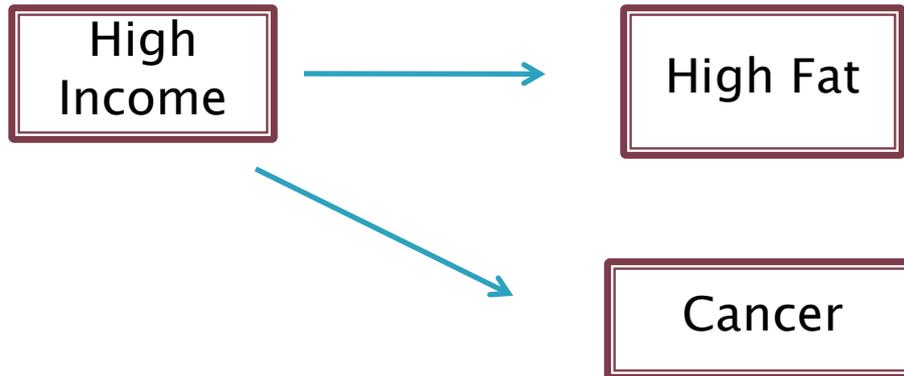


Source: K. Carroll, "Experimental evidence of dietary factors and hormone-dependent cancers," *Cancer Research* vol. 35 (1975) p. 3379. Copyright by *Cancer Research*. Reproduced by permission.

- ▶ Regression seems to suggest



- ▶ More likely to be



# Randomized Experiment

- ▶ Help separate correlation and causation.