Elementary Inference: a review

This course relies on familiarity with *frequentist* estimation and statistical inference (i.e., what you typically encounter in introductory statistics courses)

Concepts:

- Sample and population, sampling as an experiment
- Sample statistics (summaries of data) as random variables
- Estimators
- Sampling distributions of estimators
- Hypotheses, test statistics, and hypothesis testing

Elementary Inference: a review

References:

Sections $1.1 - 1.13$ of SPRM provide a review of relevant material.

Some additional sources:

- *Fundamentals of Biostatistics* by Rosner, other citations in the chapter
- *OpenIntro Statistics:* https://www.openintro.org/

Sample and Population

- A population is a group (of units or subjects) that we are interested in studying.
  - May be clearly defined - all individuals in Chicago
  - May be more abstract in terms of sampling frame (e.g., all cancer patients with specific features)
  - Sometimes very large, not easily enumerable (infinite)
- A random sample is a selection of some members of the population, s.t each member is chosen independently from the others with certain probabilities. Some questions:
  - What is target population (even if implied)?
  - Can the sampling method be expected to *represent* this target accurately ?
  - Will the sampling method *over-sample* or *under-sample* certain segments of the target population?

Statistical Inference

- Assuming we have a suitable sample, we will use it to learn about the population of interest
- **Statistical Inference**: the process of using sample data to learn about the population, typically certain population quantities

Example: We are interested the average height of college students. Here is a sample (an old one):

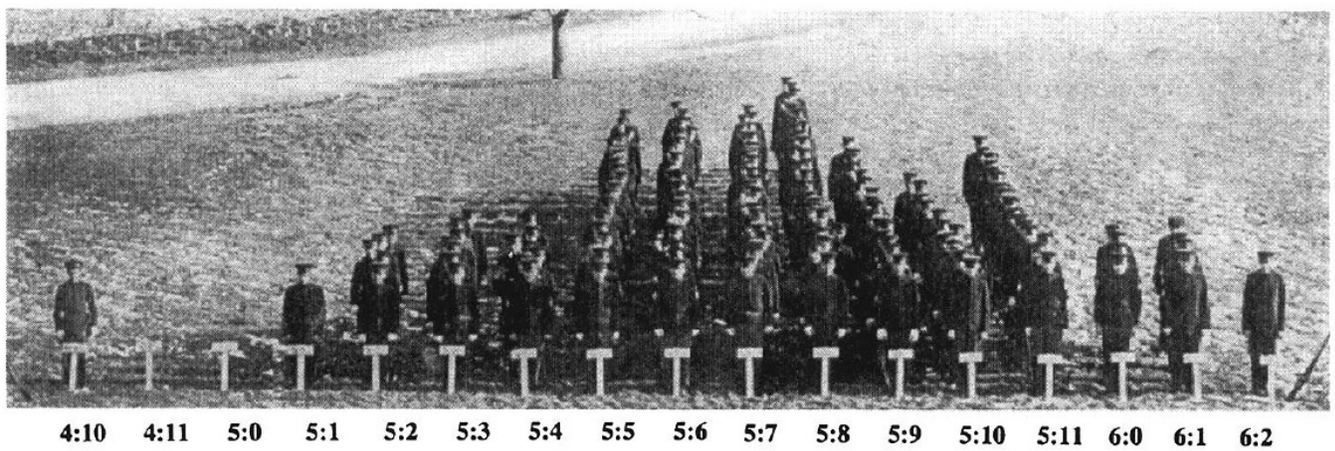


Figure 12. Living histogram of 175 male college students (Blakeslee 1914).

and a more recent one, this is apparently a tradition:

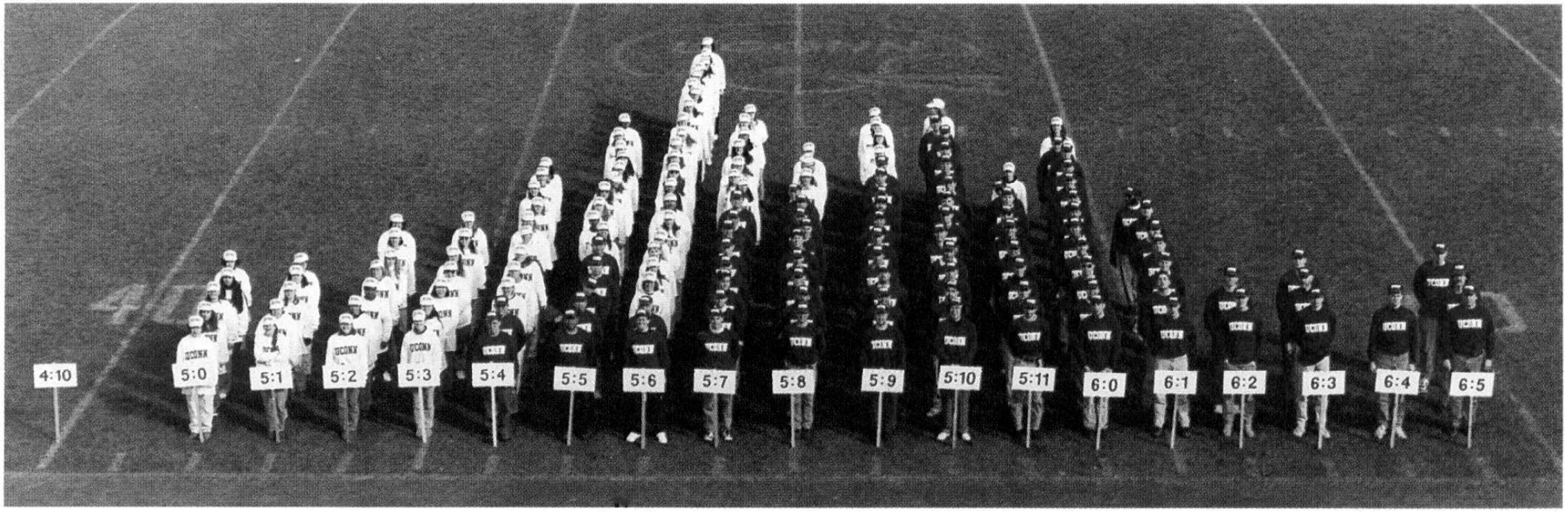


Figure 7. Living histogram of 143 student heights at University of Connecticut.

Sample Statistics

- Sampling as an experiment
  - Random samples are collections of draws from the target pop'n. Each time we do this sampling, we get a different set of draws. The quantities collected are random variables.
  - The theory of statistics provides a mathematically-defensible way to deal with sampling variation and use sample information (from one sample) to make inference about the population.
- Sample statistics as random variables
  - Numerical summaries of the samples are called sample *statistics*, The sample mean $\bar{X}$ is an example of these:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

  - Sample statistics are random variables. If we draw 1000 samples, we will have 1000 different sample means.

Sampling Distributions

- When computed from a random sample, these *statistics* are random variables and hence have probability distributions, called sampling distributions.
- It is known what sampling distributions of some key statistic ought to look like based on analytic derivations. For example, the sample mean or average:
  - If $X \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n)$.
  - Central Limit Theorem: For populations that aren't normally distributed, if $n \geq 25$, $\bar{X}$ $\text{approx.} \sim N(\mu, \sigma^2/n)$.
  - This sampling distribution is employed in making determinations about the population mean via our sample mean

Test Statistics

- The sample mean is a statistic, and also an *estimator* of a useful quantify we are interested in - the population mean μ. There are other computed quantities of the sample (e.g., statistics) that are used solely to make inferential decisions about what the population parameter of interest.
- Test statistics:
  - Are sample statistics whose sampling distributions are known (usually derived analytically), and which are used to make inference in hypothesis tests.
  - When we examine the observed test statistics in reference to its sampling distribution under some assumption about the originating population, we can assess how likely our statistic (and by extension our data) is indeed likely to have come from a population that meets that assumption.

A Statistical Hypothesis Testing Example

- Example : We are interested in comparing the average birth weights of babies born in 1995 and those born in 2000. We take the birth weight records of 40 babies each from 1995 and 2000. The sample mean (in grams) of 1995-babies is 2995 and that for 2000-babies is 3026. The sample standard deviation are 512 and 534 gms, respectively.

Q: Are the two average birth weights in the two different year periods different?

Hypothesis Testing

- **Writing Statistical Hypotheses**

  We use the conventional notation for specifying a *null* and *alternative* hypothesis:

  One-sample case, testing whether mean equals some specified μ_0

  $H_0 : \mu = \mu_0$ vs. (most generally)

  $H_a : \mu \neq \mu_0$

  OR - in the two-sample case, comparing two populations

  $H_0 : \mu_1 = \mu_2$ vs.

  $H_a : \mu_1 \neq \mu_2$

Hypothesis Testing

- in our birthweight problem

  The null hypothesis is $H_0 : \mu_{1995} = \mu_{2000}$.

  The two sided alternative hypothesis is $H_a : \mu_{1995} \neq \mu_{2000}$.

Note: We write the hypotheses in terms of the population parameters (ex. μ), not the sample quantities (ex. $\bar{X}$). The sample quantity is a statistic, the value of which is known. The population parameter is unknown and we are speculating on its likely value based on our sample.

Hypothesis Testing

- To test the hypothesis, we need to determine which test statistic to use and its sampling distribution under the null hypothesis
- In our birth weight example, the sample size is 40 for each sample. While we could use a Z (Normal distribution based) test, we might also opt for a test based on the t distribution
  - two-sample test assuming either equal variance (reasonable in this case) or unequal variance (most general test)).

Hypothesis Testing

- The estimated difference in mean weight is $\bar{X}_{2000} - \bar{X}_{1995} = 3026 - 2995 = 31$.
- The standard error of this difference (see basic stat. text) can be estimated as:

$$se(\bar{X}_d) = \sqrt{s^2_{1995}/n_1 + s^2_{2000}/n_2} = 116.97$$

- The test statistic is

$$Z = \frac{\bar{X}_d - 0}{se(\bar{X}_d)} = \frac{31}{116.97} = 0.265$$

This test statistic is *approximately* distributed as N(0,1) (Standard Normal

Critical Values and p-Values

- Now we consider the observed test statistic in reference to its sampling distribution under the null, $N(0, 1)$. Two approaches:
  - **Critical values (old way, BC):** value (from the distribution) that beyond which the probability falls below some 'small' threshold such as 0.05 (Z=1.96). Statistic from your data is compared to to this threshold. Observed statistics beyond the critical value are unlikely to have arisen from data compatible with the null hypothesis.

  - **p-values (new way, computer oriented):** Use the distribution to obtain the probability of observing a statistic value as more extreme than the one obtained, assuming that (that is, conditional on) the null being true. Expressed as $\Pr(Z > z|H_0)$

- Thus, for the second approach, if we are using the conventional 0.05 significance criterion:
  - If $p < 0.05$, we reject the null hypothesis. If the null hypothesis is true, it would be unlikely to observe such a test statistic from the data.
  - If $p \geq 0.05$, we do not reject the null hypothesis. We say the null hypothesis is *not inconsistent* with the data.
    - Note P-value ≥ 0.05 is not a proof of the null hypothesis.

- The number 0.05 is when designated a priori is called significance level or, in study design, the type I error rate, denoted by α. If the truth is no difference between mean weight in the two groups, and the same test procedure is repeated infinite times, in 5% of the tests the null will be rejected.

The Logic of p-values and Statistical Decision Making

- When we get a large statistic (small p-value), then either
  - The null hypothesis is not true - reject it
  - Something unusual happened - we obtained the statistic we did despite the null being true - called a type I error
- Since we only have a data *sample* (not the population, or multiple samples), we don't *know* when a type I error occurs, we conclude the former when we get an extreme statistic. How often is this a wrong decision? α% of instances
- In our example, based on the observed statistic, which is not at all unusual under the null hypothesis, the conclusion is the babies do not seem to be larger in year 2000.

$$\text{P-value} = P(|Z| > 0.265) = 2(1 - \Phi(0.265)) = 0.791$$

We decide that the null hypothesis is most plausible given the data

The Logic of p-values and Statistical Decision Making

Sections 1.11 and 1.13 provide a thoughtful summary of contemporary thinking about p-values and statistical significance.

- Guidelines for p-value range given (Rosner). This is not universally agreed upon nomenclature. Many feel that baseline criterion (e.g., 0.05) should simply be made smaller.
- Emphasis should be on material effects, not just significance. Effect estimate with confidence interval must be primary focus.
- Increasingly, journals have reporting standards that reinforce appropriate use of hypothesis testing, including de-emphasis on p-values, control of multiplicity, *a priori* analysis plan, etc
- In modeling, some problems are exploratory, others seek to succinctly explain a phenomenon, while still others seek to predict accurately. Error (ie false positive findings) control may differ across problems

CHECKLIST: Concepts from Elementary Statistics

Before proceeding with the materials in this course, make sure you are familiar with the following:

- Type 1 and Type 2 error rate
- Significance level, statistical power of a test
- Confidence interval
- P-value
- Z-test - Gaussian (Normal) distribution
- One-sample and paired t-test
- Two-sample t-test with equal variance assumption
- χ^2 test. F-test
- Binomial and Poisson distribution
- Hypothesis testing and estimation notation

Hypothesis Testing in Regression Analysis

In regression analysis, we use test statistics that have mostly familiar forms, as well as some new statistics, to evaluate

- Whether there is a linear association between two random variables - correlation
- Whether fixed values of one variable is associated with the value of another variable - slope in simple regression
- Whether a variable is important as a predictor in the presence of other predictors - coefficient in multiple regression
- If multiple predictors act synergistically or antagonistically - interaction effects
- The general worth of a model with respect to ability to explain variability in outcome - r^2, ANOVA F Test
- Characteristics of the estimated model that act as diagnostics

A Note about Notation in Estimation (and specifying hypotheses to be tested) in Regression Analysis

Hypothesis testing notation is arcane relative to common language. The important part is to distinguish between *parameters* and *estimates*

- **Parameters** pertain to true values in the population that we don't observe directly. Denote with Greek letters like μ (mean), ρ (correlation), and β ('slope'). We write hypotheses in terms of candidate values for parameters.
- **Estimates** are functions of the data. These are numbers with known value after estimation. These are denoted by latin-based alphabet letters with other notation ($\bar{X}$ for the mean) or by adding a carat to the Greek letter ($\hat{\mu}$). We evaluate whether the assumed population value is in effect based on our estimate.

- Test statistics (like a chi-square or t value) are also numbers, and may be one and the same w/estimators or different (usually are), but there is always a connection between them.
- In this course, we will be estimating predictor effect ('slope' $\hat{\beta}$) and testing whether the data support the notion that this estimate supports the notion that some true β is in effect

Overview of Regression Analysis

Considering a Functional vs. Statistical Model:

- We know that some variables are related in nature. For some 'output' variable Y and input X, the functional relationship may be expressed in this general form:

$$Y = f(X).$$

This is a functional model that perfectly relates Y to X (or at least so, in theory). For example, consider an equation from the CRC handbook that describes some physical law.

But oftentimes, either

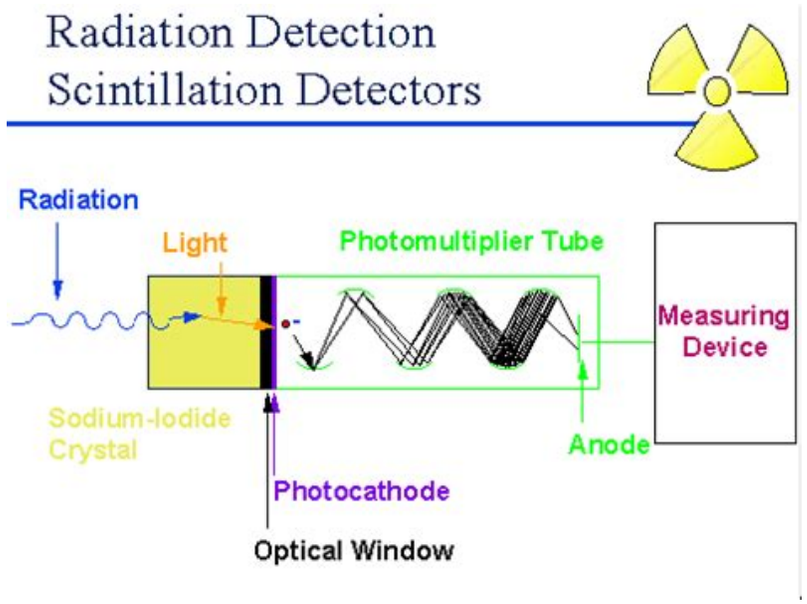
a) we do not know the functional form of the relationship and we need to approximate it by an estimated function; this is empirical, based on our limited observations, but if we have representative sampled data, it should be representative of the relationship in the population,

OR

b) Despite there being a physical 'law' governing the relationship, we wish to empirically and objectively estimate it for certain purposes

Example

Problem: Calibrate a sodium iodide crystal counter to measure radon exposure in individuals who have high radon levels in their homes. Exposure is inferred via gamma rays coming off the individual. Various factors affect the expected gamma ray emission count including geometry of the measurement setup, humidity and other atmospheric factors, 'error', etc



Example (cont)

Two ways to approach modeling the gamma ray count:

1. Physics way: Apply (nonlinear) equation describing attenuation of signal from 'ideal' and compare to measurements obtained.

$$C_{obs} = f(C_{ideal}, X_1, X_2, X_3, \ldots). = C_{ideal} X_1^{\phi_1} X_2^{\phi_2} \ldots$$

2. Fit a regression model, via an equation of the following form:

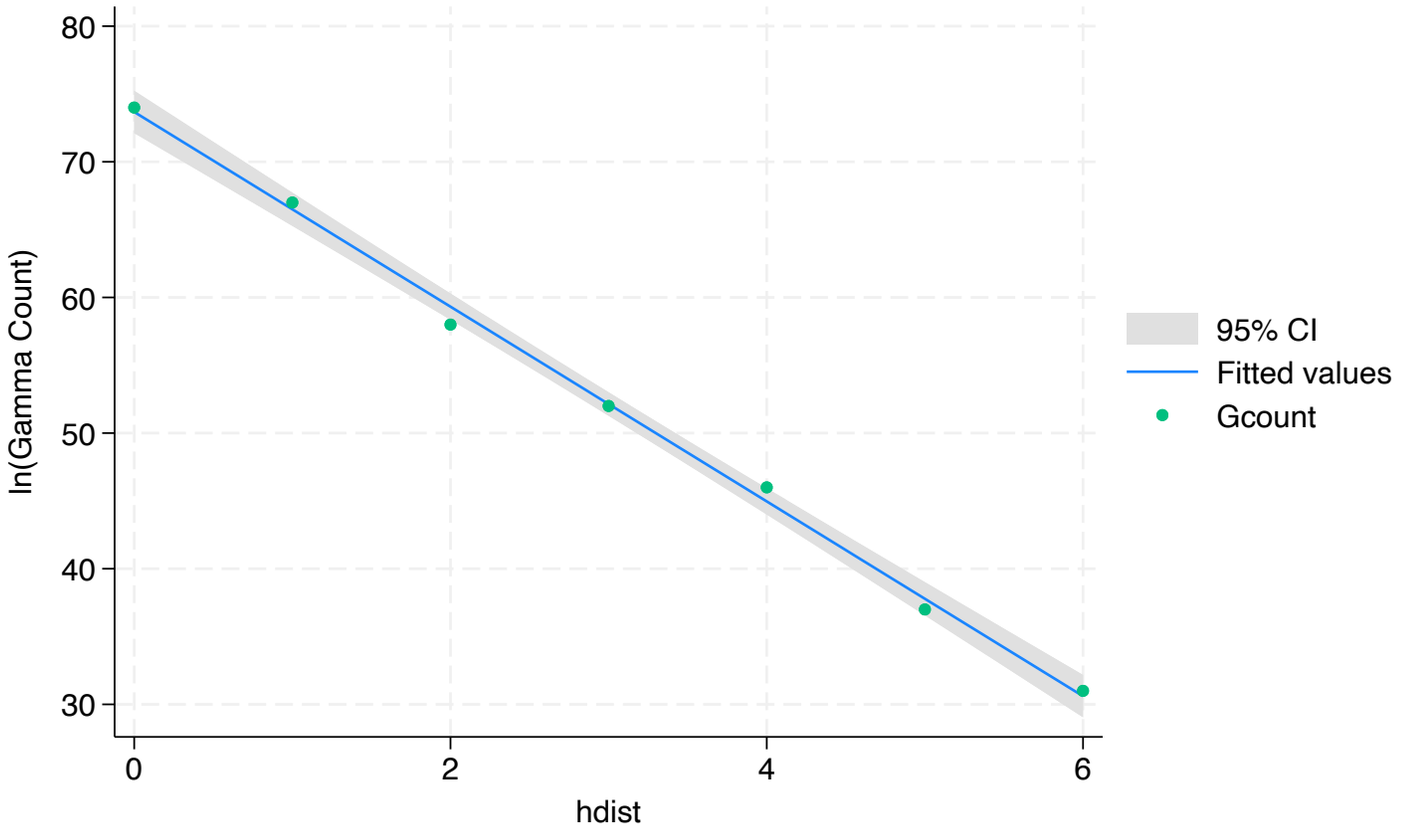
$$\log(C_{obs}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon,$$

Gamma Count Example (cont)

Place a radioactive source under the counter (ideal position to measure the 'true' gamma emission). Then move it out horizontally and take more measurements. How does horizontal position relate to the count obtained?

Gamma Count Example (cont)

Result: horizontal position is strongly related to log of counts realized



Overview of Regression Analysis

- We start with data on *response variable* Y and *predictor variables* $(X_1, X_2, \ldots, X_p)$

  For the general function

  $$Y = f(X_1, X_2, \ldots, X_p).$$

  We might opt for a linear relationship (on some scale)

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p$$

  This is still perfect prediction; instead what we have (via estimation from data) is

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_p X_{pi} + \epsilon_i,$$

  where ϵ (the"error term") denotes the discrepancy between the predicted Y value according to the specified function and the observed Y value. We review this notation formally in Ch. 2

Conducting Analyses (and Regression Analyses)

Section 1.14 of SPRM describes study design/data management issues. These are important aspects of regression analysis, and we will discuss in the context of specific data sets and analyses. Please review this section and 1.15 (causality) on your own.

This book/course is predominantly about linear regression. One might ask, why *linear* regression? Seems restrictive and simplistic:

- Few processes are actually linear?
- Forces monotonic relationship(s) between X and Y, no 'threshold' effects, etc

However, keep in mind that

- Model is necessarily and intentionally simple, seeks *data reduction*
- Model says response can be written as a linear function of factors; these factors can be nonlinear functions of X: x^2, $\log(x)$, etc.

Scope of Regression Analysis Problems

Statisticians are sometimes said to reduce everything to linear forms and draw lines through points. In this course we will go further, specifically to

- Draw lines through points! We will use linear regression models to find linear relations between a continuous outcome measure and predictor variables.
- Model/predict probabilities, using a seemingly indirect approach that relates binary outcomes (yes/no, 1/0, etc) to predictor variables .
- Model counts of events, such as disease incidence, relating the count or rate of events to predictor variables.

All of these can be approached as a linear model on some scale

Conducting Regression Analyses

Specifying, executing, interpreting (fitting model, critiquing, revisiting objectives) will be a major focus of this course:

- There are many methods to critique, check linear model suitability; Proper regression analysis will use these extensively
- Emphasis in regression is often on estimating effects rather than hypothesis testing, latter is used to confirm veracity of former
- Alternatives exist and should be used where indicated - do not force linear model when inappropriate to use
- George Box: "essentially, all models are wrong, but some are useful"

A clearly important linear regression problem:

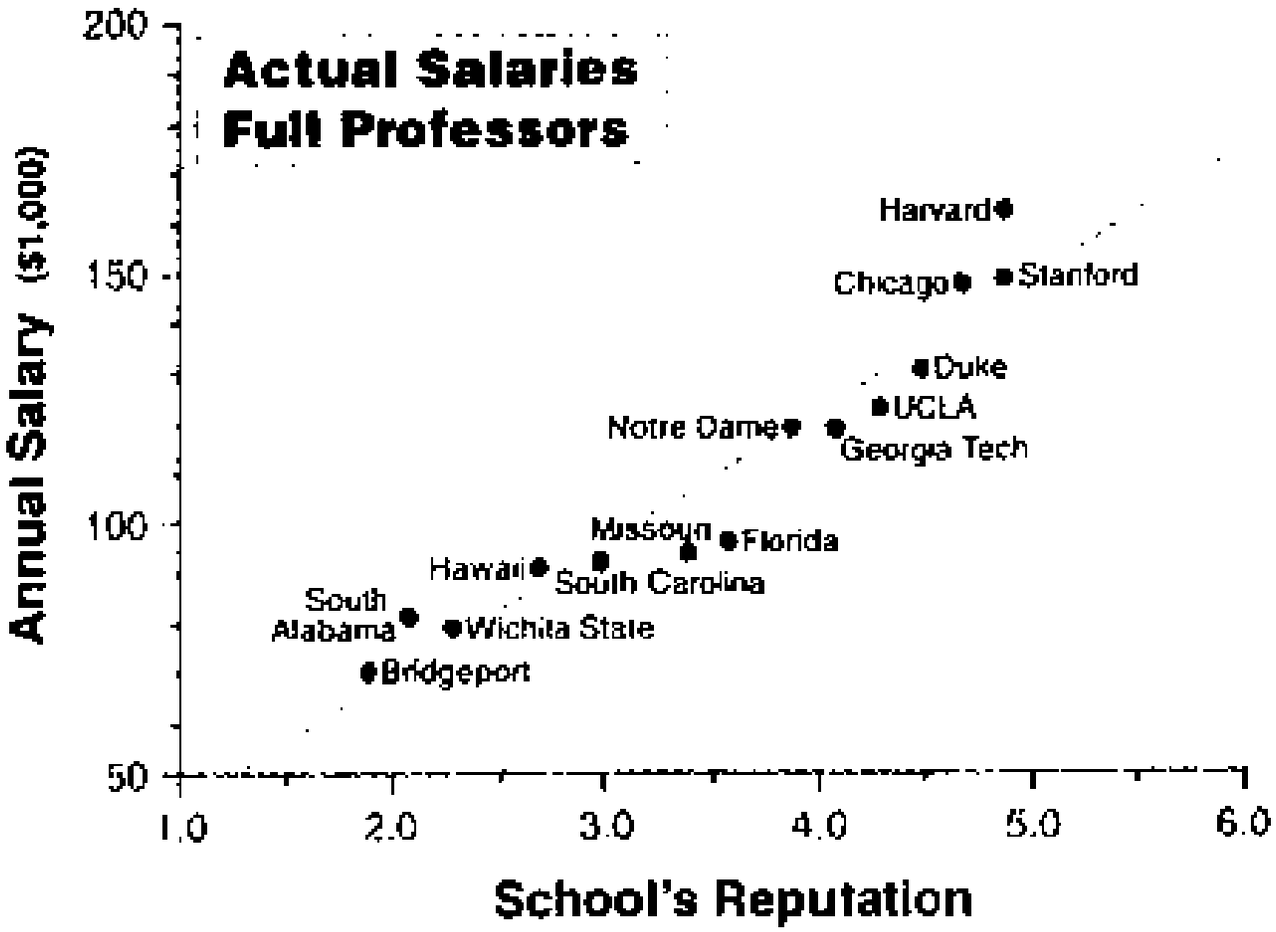


FIGURE 1. Mean 9-month salaries of full professors as a function of the institution's academic reputation. Academic year 2004-05. The correlation coefficient for this sample of 14 universities is $r = 0.96$ ($p < 0.001$). Salary data from the AAUP's annual report.[2] Reputation data from the *U.S. News & World Report*'s college guide (peer assessment).[3]

Conducting Regression Analyses

Here, we might proceed as follows:

- What is the hypothesized relationship (directionally, not necessarily in terms of magnitude)?
- Is a linear association reasonable (on some scale)?
- What are additional issues that should be considered (such as the sample source and range, etc)?
- Are there additional factors to consider before drawing a conclusion?