Stat 430: EMPIRICAL METHODS FOR COMPUTER SCIENCE RESEARCH

Karin S. Dorman

Department of Statistics
Iowa State University

August 25, 2009

The Scientific Method

1. Pose questions.
2. Formulate a hypothesis (**probability modeling**).
3. Design an experiment and collect the results (**experimental design**).
4. Interpret data, answer questions, return to step 1 (**hypothesis testing, estimation**).

The Question(s)

1. Which of you, the students in Stat 430, are prepared for the course?
2. Are there differences in the preparedness of students on different topics?
3. Is there a change in the preparedness of students compared to last year?

The Question(s)

1. Which of you, the students in Stat 430, are prepared for the course?
2. **Are there differences in the preparedness of students on different topics?**
3. Is there a change in the preparedness of students compared to last year?

Data Collection

1. Do you know what is the definition of conditional probability?
2. Do you know what is *Bayes' Rule*?

The Data

Topic	Count	Percentage
Conditional probability	12	0.44
Bayes' Rule	10	0.37

Answering the Question

Is there a difference in the students' knowledge of conditional probability and Bayes' rule?

If I am asking with respect to this class only, then my answer is immediate.

However, if I want to make a general statement about the population of students that could enroll in Stat 430 ever, then I need to make an *inference*. I utilize *a model* and the *sample* (you!) to draw a conclusion.

It is very important to identify the *population* to which the conclusions should generalize. This class or all potential students?

State your questions precisely.

The (Precise) Question

Do students in the population of all students who want to take Stat 430 differ in their knowledge of conditional probability and Bayes' rule?

Knowing that I want to answer the question for the hypothetical population of all students interested in Stat 430, I recognize that my data is incomplete. It represents only a *sample* of the full population. I may not be able to identify all students who want to take Stat 430, and I certainly cannot question future students, so there is no way I can answer the question directly. I need to formulate a more careful *model* of reality in order to draw any conclusions.

The Hypothesis

Model. I hypothesize that you all represent a *random sample* from a hypothetical *population* of students that want(ed) to take Stat 430 in the past, now, or in the future. Further, I assume

- students answer questions independently,
- students are identical (no change due to date, background, etc).

Error. Errors are made at every stage of the process. There are undoubtedly errors in my model. Also, in collecting your answers, error is introduced. I may miscount. You may not be sure how to answer a question. You may not respond truthfully.

Identify errors in your model and data collection process.

Answering the Question (Inference)

Do students in the population of all students who want to take Stat 430 differ in their knowledge of conditional probability and Bayes' rule?

Let p_c be the probability a random student in the population knows conditional probability. Let p_b be the probability a random student in the population knows Bayes' rule. p_c and p_b are model (population) parameters. My specific hypothesis related to this question is:

$$H_0 : p_c = p_b$$

Notice, the hypothesis is a statement about the population, *not* my sample.

More Complicated Questions

What characteristics of a Stat 430 student predicts success in the course?

For example, suppose x_c and x_b indicate whether the student claimed to know conditional probability and Bayes' rule before entering the class. Then, I might hypothesize that success

$$y = b_0 + b_c x_c + b_b x_b + \epsilon$$

where y is the final percentage in the course, b_0, b_c, b_b are model parameters, and $\epsilon \sim \text{Normal}(0, \sigma^2)$.

A specific hypothesis addressed toward the question would be

$$H_0 : b_c = 0$$

Probability Review Outline

- Probability experiment with an outcome in a discrete or continuous sample space Ω.
- Set operations: union, intersection, mutually exclusive, exhaustive, complement, empty set, distributivity
- Probability: Kolmogorov's Axioms, addition rule, multiplication rule, counting methods* (summation, multiplication principles, ordered, unordered, with replacement, without replacement), law of total probability*, conditional probability*, Bayes' rule*, independence*
- Random variables: cdf, pmf, pdf
  - Discrete: Uniform, Bernoulli, Binomial, Geometric, Poisson
  - Continuous: Uniform, Exponential, Gamma, Erlang
  - Expectation*, moments, variance, Chebyshev Inequality
  - Change of variable:
- Multiple random variables: joint distributions, marginal distributions, independence, expectation, variance, covariance, correlation, conditional distribution, conditional expectation, change of variable,