Applied Side of Statistics – Some Real Case Studies

Sat Gupta
University of North Carolina -Greensboro
Professor of Statistics
http://www.uncg.edu/mat/faculty/sngupta/
sngupta@uncg.edu

December 11, 2015
Tata Institute of Fundamental Research
Mumbai, India

Outline

Basics

- Statistics Vs Mathematics
- Statistical Studies
- Statistical Significance

Some Applications

- Data Confidentiality and Respondent Privacy
- RRT Models – Warner & Greenberg Models
- Statistical Consulting Based Case Studies

Statistics Vs Mathematics

- In mathematics, conclusions are driven by precise models. In statistics, conclusions are often data driven

- In mathematics, we make statements with no room for error. In statistics, there is always an error element

- In mathematics, we make statements like X>Y. In statistics, the corresponding statement typically is ***X*** **is significantly greater than *Y***

Statistical Studies

- **Descriptive Studies:** We collect some data from a target population and report data characteristics. We may be able to speculate on some trends
- **Inferential Studies:** We collect some data **randomly** from a target population, analyze the data, and make inference **(extrapolation)** about the target population

Example: How to decide if a large lot of diagnostic test kits meets specifications and should be released for sale

- **Observational Studies:** You analyze data that was not collected randomly. No inference can be made in such cases but trends may be detected. Causation can’t be claimed
- **Experimental Studies:** Observational units are randomly assigned to different groups. Data analysis can lead to causational claims

Statistical Significance

A finding is called statistically significant if the chance of seeing that finding, or even more extreme finding, just by chance, is less than .05

Stochastic Scrambling and Descrambling

- At the front end, data may be of sensitive nature and there is need to circumvent social desirability bias by offering respondents privacy
- At the back end, data confidentiality need to be maintained since the researcher is bound by a confidentiality clause
- Descrambling can happen only at aggregate level and not at individual level
- Synthetic data may be used

Randomized Response Models

- Scientific alternative to **Bogus or Pseudoscientific** methods
- Have been around since 1965
- Many new variations of RRT models have come about in the past 50 years

Warner’s Model (1965) – The Original Model

- *л* = Proportion of a sensitive characteristic in a population
- *p* = Proportion of cards in a deck asking sensitive question **directly**
- *1- p* = Proportion of cards in the deck asking the sensitive question **indirectly**
- p_y = Probability of Yes response
- $p_y = \boldsymbol{p}\text{л} + (\mathbf{1} - \boldsymbol{p})(\mathbf{1} - \text{л})$

Parameter Estimation

- $\pi = \frac{p_y-(1-p)}{2p-1}, \; p \neq 1/2$

- $\hat{\pi} = \frac{\frac{n_1}{n}-(1-p)}{2p-1}$

- n_1= Number of "Yes" responses in the sample

Greenberg Unrelated Question Model (1969)

- π = Prevalence of a sensitive characteristic
- π_u = Prevalence of some innocuous characteristic
- p = Proportion of cards in the deck with sensitive question
- $1 - p$ = Proportion of cards in the deck with the innocuous question
- $p_y = \boldsymbol{p}\text{л} + (\mathbf{1} - \boldsymbol{p})\boldsymbol{\pi}_{\boldsymbol{u}}$

- $\pi = \frac{p_y - (1-p)\pi_u}{p}$
- $\hat{\pi} = \frac{\frac{n_1}{n} - (1-p)\pi_u}{p}$

Gupta et al. (2013) – Optional Models

- The probability of a 'yes' response P_Y is given by:

$$P_Y = (1\text{-}W)\pi + W[p\pi + (1\text{-}p)\pi_u]$$

- Use Split Sample Approach

- The probability of 'yes' response in the i^{th} (i =1,2) sub-sample is given by:

$$P_{Yi} = (1\text{-}W)\,\pi + W[p_i\,\pi + (1\text{-}p_i\,)\pi_u]$$

$$\pi_x = \frac{P_{Y_1} - \lambda P_{Y_2}}{1-\lambda}\text{, where}\quad \lambda = \frac{1-p_1}{1-p_2}$$

Recent Applications of RRT

Ostapczuk, Martin, Jochen Musch, and Morten Moshagen (2009):

A randomized-response investigation of the education effect in attitudes towards foreigners, *European Journal of Social Psychology*, 39 (6)

Spears- Gill, Tracy., Tuck, Anna., Gupta, Sat., Crowe, Mary., Jennifer Figuerova (2013):

A Field Test of Optional Unrelated Question Randomized Response Models – Estimates of Risky Sexual Behaviors, *Springer Proceedings in Mathematics and Statistics*, Vol. 64, 135-146

Education Effect in Attitudes Towards Foreigners in Germany

- Under direct questioning conditions, 75% of the highly educated expressed xenophile attitudes, as opposed to only 55% of the less educated.

- Under randomized-response conditions, 53% xenophiles among the highly educated, and 24% among the less educated

Spears-Gill et al. (2013) - Field Test: Estimates of Risky Sexual Behaviors

Sensitive question

Have you ever been told by a healthcare professional that you have a sexually transmitted disease(STD)?

Unrelated question

Were you born between January 1st and October 31st?

- Target population: undergraduate students enrolled at UNCG during the 2012-13 academic year.
- Sample: 878 subjects from undergraduate level class sections in math and statistics

- Methods:

  optional unrelated question RRT

  check-box survey method

  direct face-to-face interview

Estimate of STD Diagnosis Prevalence

Method	$\hat{\pi}_X$	**95% CI***
Optional RRT	0.0367	(0.0159, 0.0576)
Check Box Method	0.0900	(0.0438, 0.1362)
Face-to-face Interview	0.0200	(-0.0042, 0.0442)

Estimates of Sensitivity Level

Question	Sensitivity Level	95% CI*
Number of Sexual Partners	0.6098	(0.5981, 0.6215
STD History	0.7730	(0.7712, 0.7748)

Zhang et al. (2015): Some Risk Predictors of STD

Set STD incidence as dependent variable;

and age, gender, and number of sex partners as independent variables.

Use logistic regression to test which variables are significant predictors of STD.

Some Significant Predictors of STD

	B	Sig.	Exp （B）
Age	0.189	0.000	1.209
Gender	3.003	0.004	20.155
Sex Partners	0.486	0.000	1.626

Results

Gender, age, and number of sex partners are all significant factors that affect STD incidence.

- Women have 20.155 times the odds men have, with all else kept equal.
- The odds of STD increase by 62% with each additional sex partner.

The Beautiful World of Statistical Consulting

Some Case Studies

FDA's Milk Testing Protocol

- Sensitivity = P (Test result is positive / sample is contaminated)
- Specificity = P (Negative Result / Sample is not contaminated)

FDA Protocol for Milk Testing Diagnostic Test Kits

Sensitivity > .90 with 95% degree of confidence

Specificity > .90 with 95% degree of confidence

Formal Test of Compliance

In a randomly selected sample of test kits, sensitivity should be greater than .90 after accounting for statistical uncertainty

- *N* = Number of contaminated samples = 20,
- *X* = Number of positive tests = 20

Sample Sensitivity = 20/20 = 100%
but is it enough?

- *P-Value* = P (X ≥ 20 / N = 20, p = .90) = .1216

- Not enough evidence that the lot meets specifications

- $N = 30, X = 30$

Sample Sensitivity = 30/30 = 100%

- *P-Value* = $P(X \geq 30 / N = 30, p = .90) = .0424$

- There is enough evidence that the lot meets specifications

Amoxicillin Data:

Concentration Level (PPB)	0	3	4	6	8	10
Positive Results	0/60	0/30	0/30	3/29	15/30	29/29

According to FDA,
the test kit does not meet specifications!

Is this conclusion valid?

Concentrating only on Level 10 PPB

- Sample Sensitivity = 29/29 = 100%
- Approximate 95% Confidence Interval using Normal Approximation cannot be calculated
- 95% Exact Lower Confidence Limit Corresponding to 29/29 Correct Results: .902

Generalized Linear Models

Logistic Regression Model:

$$\ln(p/(1-p)) = \alpha + \beta x$$

Probit Regression Model:

$$\phi^{-1}(p) = \alpha + \beta x$$

Gompertz Regression Model:

$$\ln(-\ln(1-p)) = \alpha + \beta x$$

95% Lower Confidence Limits on Sensitivity

Logistic Regression Model:	.8975	
Probit Model:	.8729	(FDA Approach)
Gompertz Model:	.9375	

Gompertz Model provides the best fit!

An Age Discrimination Lawsuit

A 62 year old engineer vs. United Technology

Total Number of Employees	183
# Employees laid-off	5
Plaintiff Age:	62 years

- Ages of others who were laid-off:

  61, 44, 53, and 45

How do you make a case that age discrimination may have occurred?

Initial Observations

- Look at the ages of two groups – those who were fired and those who were not fired. Argue that the average age of those who were fired was significantly higher than the average age of those who were not fired.

Average Age of those laid-off	53.00
Average age of those not laid-off	45.37

Statistical Analysis

You can use a *t-test* or *Wilcoxon Rank Sum Test.*

Two-sample t-test p-value:

.020 (2-sided)

Wilcoxon Rank Sum Test p-value:

.058 (2-sided)

The defense made the fatal error of reporting a two-sided p-value of .058 for the *Wilcoxon Rank Sum Test.*

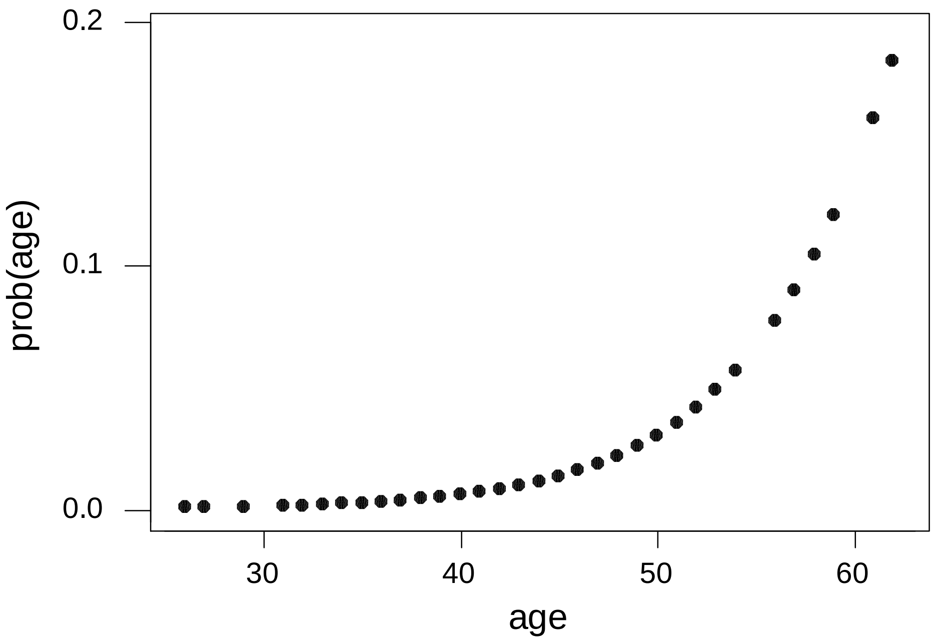
Binary Logistic Regression

Use the *Logistic Regression Approach* and argue that age is a significant predictor after controlling for other important indicators.

Age *p-value*	.028

Estimated lay-off probability for various ages

Figure 1:Probability of termination for different ages



Medicare/Medicaid Overpayments

A specific overpayment case:

- Total number of claims: 540 = 460 + 80
- Total Paid for these claims: $28,809.00
- Sample Size: 50 + 50 = 100
- Each of the sampled claims is audited
- For any claim:
  Overpayment = Paid Amount – Audit Amount

Overpayment Calculations

Use Stratified Random sampling

- Point Estimate of Total Overpayment: $ 6,917

- 90% Confidence Interval for over payment amount for the population based on the 100 sampled claims only: (5,006, 8,827)

- $ 5006 is the 95% lower confidence limit

Why Trust this calculation?

- 95% Confidence Interval for total paid amount for the population based on the 100 sampled claims only:

  (27,896, 29,313)

- Actual Total Payment: $28,809.00
- Sample data does lead to good population estimate!

Thank You!

Questions/Comments