Name-Brand vs. Off-Brand: A Twist on Taste Testing for a Mathematical Statistics Course

Eric M. Reyes

Rose-Hulman Institute of Technology

JSM 2014

Outline

Dataset and Story

The Mathematical Statistics Course

- Students have a strong mathematics background including: Calculus, Probability, Programming, and Introductory Statistics, (possibly) Analysis.

The Mathematical Statistics Course

- Students have a strong mathematics background including: Calculus, Probability, Programming, and Introductory Statistics, (possibly) Analysis.
- Topics include probability, properties of random samples, estimation, and inference via maximum likelihood.

The Mathematical Statistics Course

- Students have a strong mathematics background including: Calculus, Probability, Programming, and Introductory Statistics, (possibly) Analysis.
- Topics include probability, properties of random samples, estimation, and inference via maximum likelihood.
- Exposes students to the theory underlying many methods encountered in other courses.

The Mathematical Statistics Course

- Students have a strong mathematics background including: Calculus, Probability, Programming, and Introductory Statistics, (possibly) Analysis.
- Topics include probability, properties of random samples, estimation, and inference via maximum likelihood.
- Exposes students to the theory underlying many methods encountered in other courses.
- It is really easy to divorce this course from the pedagogical tools we use in an Introductory Statistics course.

Background for Students

Name-Brand vs. Off-Brand: Can We *Really* Distinguish Between the Two?

Activity and Assignment

Data Collection

Question: What is the probability that a consumer can *actually* discern the difference between Cheerios and an off-brand version?

Data Collection

Question: What is the probability that a consumer can *actually* discern the difference between Cheerios and an off-brand version?

Taste Test:

- Participants were instructors at the university.

Data Collection

Question: What is the probability that a consumer can *actually* discern the difference between Cheerios and an off-brand version?

Taste Test:

- Participants were instructors at the university.
- Double-blind.

Class Discussion:

- What is the benefit of blinding and how might it be implemented?

Data Collection

Question: What is the probability that a consumer can *actually* discern the difference between Cheerios and an off-brand version?

Taste Test:

- Participants were instructors at the university.
- Double-blind.
- Each participant was given $m = 6$ samples (3 name-brand and 3 off-brand in randomized order).

Class Discussion:

- What is the benefit of blinding and how might it be implemented?
- How will randomization be utilized in this study?

Data Collection

Question: What is the probability that a consumer can *actually* discern the difference between Cheerios and an off-brand version?

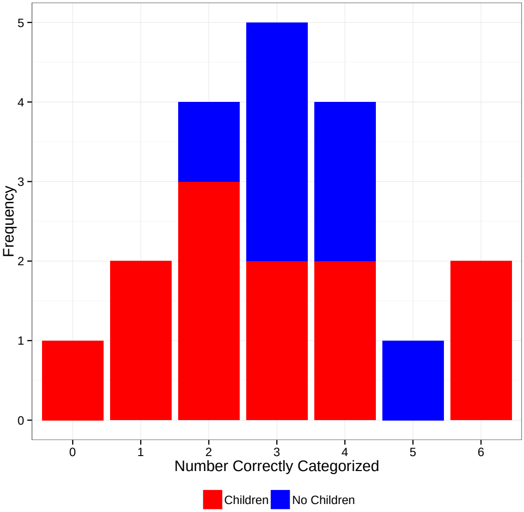
Taste Test:

- Participants were instructors at the university.
- Double-blind.
- Each participant was given $m = 6$ samples (3 name-brand and 3 off-brand in randomized order).

Class Discussion:

- What is the benefit of blinding and how might it be implemented?
- How will randomization be utilized in this study?
- What variables should be recorded in this study?

The Data


5
4
3
2
1
0
Frequency
0
1
2
3
4
5
6
Number Correctly Categorized
Children
No Children

The Assignment

Question: What is the probability that a consumer can *actually* discern the difference between Cheerios and an off-brand version?

- Describe the distribution of the probability of actually discerning between the name-brand and off-brand products.

The Assignment

Question: What is the probability that a consumer can *actually* discern the difference between Cheerios and an off-brand version?

- Describe the distribution of the probability of actually discerning between the name-brand and off-brand products.
- Account for the fact that a correct response could be due to a true discernment or a lucky guess.

The Assignment

Question: What is the probability that a consumer can *actually* discern the difference between Cheerios and an off-brand version?

- Describe the distribution of the probability of actually discerning between the name-brand and off-brand products.
- Account for the fact that a correct response could be due to a true discernment or a lucky guess.
- Construct a poster-presentation summarizing your analysis and results.

Analysis

Model for Data-Generating Process

Model Proposed by Morrison [*Am. Stat.* (1978)]

- P_i is the probability the i-th subject can actually discern between the name-brand and off-brand.

Model for Data-Generating Process

Model Proposed by Morrison [*Am. Stat.* (1978)]

- P_i is the probability the i-th subject can actually discern between the name-brand and off-brand.
- Then, the probability of correctly categorizing the response is

$$C_i = P_i + \frac{1}{2}(1 - P_i)$$

Model for Data-Generating Process

Model Proposed by Morrison [*Am. Stat.* (1978)]

- P_i is the probability the i-th subject can actually discern between the name-brand and off-brand.
- Then, the probability of correctly categorizing the response is

$$C_i = P_i + \frac{1}{2}(1 - P_i)$$

- Assume $P_i \overset{IID}{\sim} Beta(\alpha, \beta)$.

Model for Data-Generating Process

Model Proposed by Morrison [*Am. Stat.* (1978)]

- P_i is the probability the i-th subject can actually discern between the name-brand and off-brand.
- Then, the probability of correctly categorizing the response is

$$C_i = P_i + \frac{1}{2}(1 - P_i)$$

- Assume $P_i \overset{IID}{\sim} Beta(\alpha, \beta)$.
- Then, the number correct is $Y_i \mid C_i \sim Bin(6, C_i)$.

Probabilistic Model

- Letting $f(y \mid \alpha, \beta)$ represent the marginal density of Y_i, then we have that

$$\ell(\alpha, \beta \mid y) = \sum_{i=1}^{n} \log\left[f\left(y_i \mid \alpha, \beta\right)\right]$$

Probabilistic Model

- Letting $f(y \mid \alpha, \beta)$ represent the marginal density of Y_i, then we have that
$$\ell(\alpha, \beta \mid y) = \sum_{i=1}^{n} \log\left[f\left(y_i \mid \alpha, \beta\right)\right]$$
- Determining the marginal density is a good exercise in its own right (Casella and Berger [Statistical Inference (pg 197)]).

Probabilistic Model

- Letting $f(y \mid \alpha, \beta)$ represent the marginal density of Y_i, then we have that

$$\ell(\alpha, \beta \mid y) = \sum_{i=1}^{n} \log\left[f\left(y_i \mid \alpha, \beta\right)\right]$$

- Determining the marginal density is a good exercise in its own right (Casella and Berger [Statistical Inference (pg 197)]).
- Students are expected to show that

$$f(y \mid \alpha, \beta) = \frac{\binom{m}{y}\Gamma(\alpha+\beta)}{2^m\Gamma(\alpha)\Gamma(\beta)} \sum_{k=0}^{y} \binom{y}{k} \frac{\Gamma(\alpha+k)\Gamma(\beta+m-y)}{\Gamma(\alpha+\beta+m-y+k)}$$

Computation of MLE's

- No closed-form solution exists for $\widehat{\alpha}$ and $\widehat{\beta}$.

Computation of MLE's

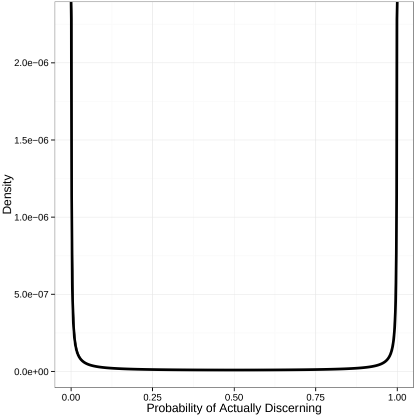
- No closed-form solution exists for $\widehat{\alpha}$ and $\widehat{\beta}$.
- Students constructed R programs to compute the estimates using the data.

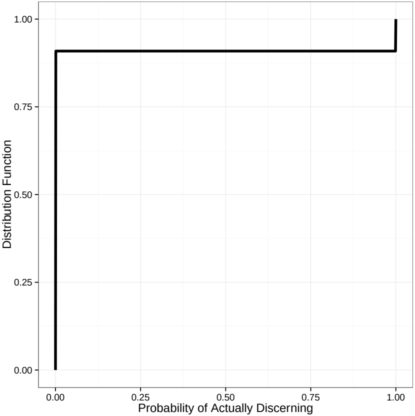
Computation of MLE's

- No closed-form solution exists for $\widehat{\alpha}$ and $\widehat{\beta}$.
- Students constructed `R` programs to compute the estimates using the data.
- Due to numerical instability, care had to be given to the optimization routine chosen and the coding of the likelihood.

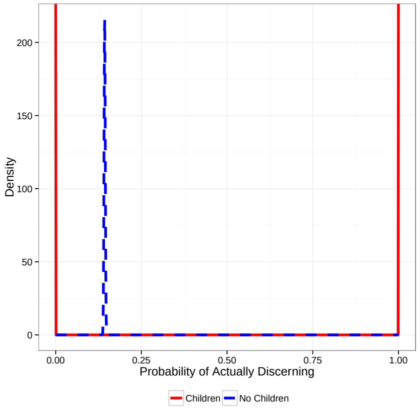
Results

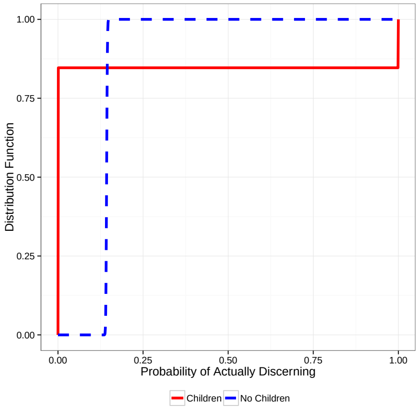
Results for Primary Question





Results Comparing Those with and without Children





Conclusions

- Students gave positive feedback on the activity, particularly seeing a problem through from design to results.

Conclusions

- Students gave positive feedback on the activity, particularly seeing a problem through from design to results.
- Activity ties together several topics used throughout the course, and is a nice capstone to maximum likelihood.

Conclusions

- Students gave positive feedback on the activity, particularly seeing a problem through from design to results.
- Activity ties together several topics used throughout the course, and is a nice capstone to maximum likelihood.
- Lends itself well to a short poster presentation.

References

- Casella G and Berger RL.
  Statistical Inference (2nd ed, pg 197).
  Pacific Grove, CA: Thomson Learning, 2002.
- Morrison DG.
  A Probability Model for Forced Binary Choices.
  The American Statistician, 32(1):23-25, 1978.

Contact Information

Eric M. Reyes
Department of Mathematics
Rose-Hulman Institute of Technology
5500 Wabash Ave.
Terre Haute, IN 47803

web: www.rose-hulman.edu/~reyesem
email: reyesem@rose-hulman.edu