# Customer Analytics

Lecture 1

Tilburg University

# Agenda

- Course administration
- What is testing and why is it useful?
- Uncertainty and probability of making a mistake; how big should the test be?
- Making smarter use of the test with targeting

# Teachers

Instructor
George Knox
G.Knox@tilburguniversity.edu

email about:
Lectures, content

Assistant
Anne van der Vliet
A.vdrVliet@tilburguniversity.edu

email about:
Assignments, administration

# Marketing: then and now

Product-centric

Transaction-focused

Customer-centric

Relationship-focused

Customers are assets that generate profits over time

# Customer lifecycle

Customer **development**: change in behavior over time: buying more (up-selling) or different things (cross-selling)

Customer **acquisition**:
how customers are "born" or first contact with the firm.

Customer **retention**: preventing customer "death" or churn.

Marketing is about acquiring, developing and retaining customers

# Customer analytics

Using customer data and statistical models to make business decisions:

- Who should be targeted for … a marketing campaign, churn prevention, cross-selling, acquisition?
- Should we do a test before we roll it out? How big?
- How many subscriptions/transactions can we predict over time for a cohort of customers?
- How valuable is a customer to the firm over his or her lifecycle? How does it differ across customers?

# Lectures 1-5: Short-term analytics

- **Testing and Uncertainty**: Why test? Quantifying uncertainty; how large should the test be?

- **Models for selecting customer to target**: which customers should be selected for e.g., acquisition, retention, direct mailing?

- **Models for customer development**: collaborative filtering, cross-selling
  - Guest lecture: Barrie Kersbergen (Bol) on Recomender systems in practice

# Lectures 6-9: Long term analytics

- How does the customer base change over time as customers drop out? Why does retention increase over time?

- **Customer lifetime value (CLV)**: who are the most valuable customers: how do you calculate the value to the firm of the customer over his or her lifecycle?
  - Guest lecture: Coolblue, Implementing CLV

# Grading

1. Individual assignments: 30%

2. Computer exam (individual): 70%

# Grading (2)

To pass the course you need:

1. final grade ≥ 6
2. exam grade ≥ 5

The assignment grade still counts if you take the resit

# Assignments

- Each lecture has an assignment
- Due a week after following Sunday
  - Late assignments not accepted
- It's OK if you discuss with others, but all assignments are to be done individually.
- Testvision software (If you have problems with part, be sure to email Anne)

# Data sets & software

- The course is organized around several data sets that illustrate an important concept.
  - All these examples will be "hands-on" and have an emphasis on real-time problem solving.

- We're using R this year (not SPSS as past years)
  - Advantages: widely used & lots of contributed software, free
  - Disadvantages: programming language, unpredictability of packages, updates

- R notebooks in the computer lab

# Readings

- Book: Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin. *Why Database Marketing?* Springer New York, 2008. **[BKN]**

- Articles: Other articles and material you can find on canvas under modules.

# Weekly schedule

- Lecture
- Computer lab
- Q&A

**Everything will be given online; no in-person/offline sessions**

**Everything will be given live and recorded**

# Testing and Uncertainty

# E-Beer

- E-Beer sells beer over the Internet and currently has about 50,000 customers
- A customer selects the favorite brand, pays, and within 1 hour the ordered amount of beer is delivered at the specified address
- To boost sales, E-Beer developed a mailing to send to their customers.
- Each mailing contains a flyer to remind customers of the offered service and a key ring with the name and web address of the company

# Campaign costs

- Each mailing costs € 1.50

- Sending it to all customers would mean total costs of

$$€\,1.50 \times 50000 = €\,75000$$

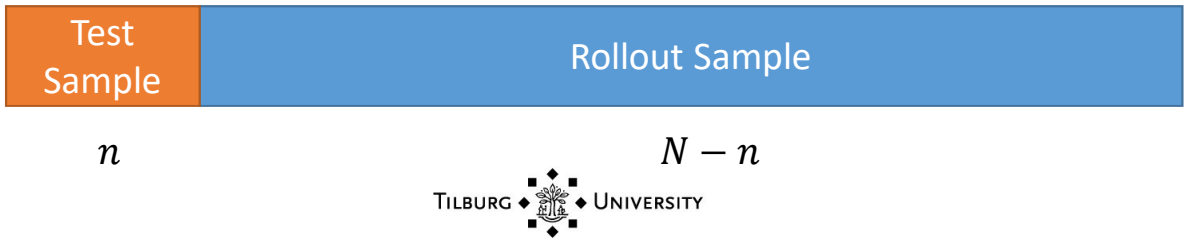
Is it worth it?  Do benefits > costs?

The problem is that the **benefit** is uncertain!

# Testing

The objective of testing is to obtain more information before committing a large amount of resources and, hence, reduce the risk of possible failure.

1. Randomly select some customers; call this test sample (size = $n$).
2. Send them mailing, collect & analyze responses
3. Use results to decide whether to send to the rest of the population (size $N - n$, rollout sample).



# Results of test

- Assume we choose a test sample of size 5000. So, we randomly select 5000 customers and send them the mailing.

- Results of test mailing
  - 175 out of 5000 respond. So the estimated response rate: $\hat{p} = 175/5000$
  - We assume the margin or profit per response is €50: $\mathrm{m} = 50$

- So should we do the rollout? How much would we expect to make if we send to the rest (rollout sample)?

# Expected rollout profits

- Assuming they are like the test sample, (which they are if randomly sampled):

| | **# customers** | **Profit per customer** |
|---|---|---|
| $\mathrm{E}[\text{rollout profit}]$ | $= (N - n)(m \cdot \hat{p} - c)$ | |
| | $= (45000)(50 \cdot 0.035 - 1.50)$ | |
| | $= \mathbf{11250}$ | |

where

$m$ is the margin (profit) per response (in euros)
$\hat{p}$ is the estimate of the response rate
$c$ is the cost of marketing

# Option value

- Therefore, because our expected rollout profit is positive, we roll it out to the rest of the sample.

- The test gives us the option – not the obligation – to rollout. We only roll out when:

$$E[\text{rollout profit}] > 0$$

$$\Rightarrow \hat{p} > \frac{c}{m}$$

in the example, what's the threshold?

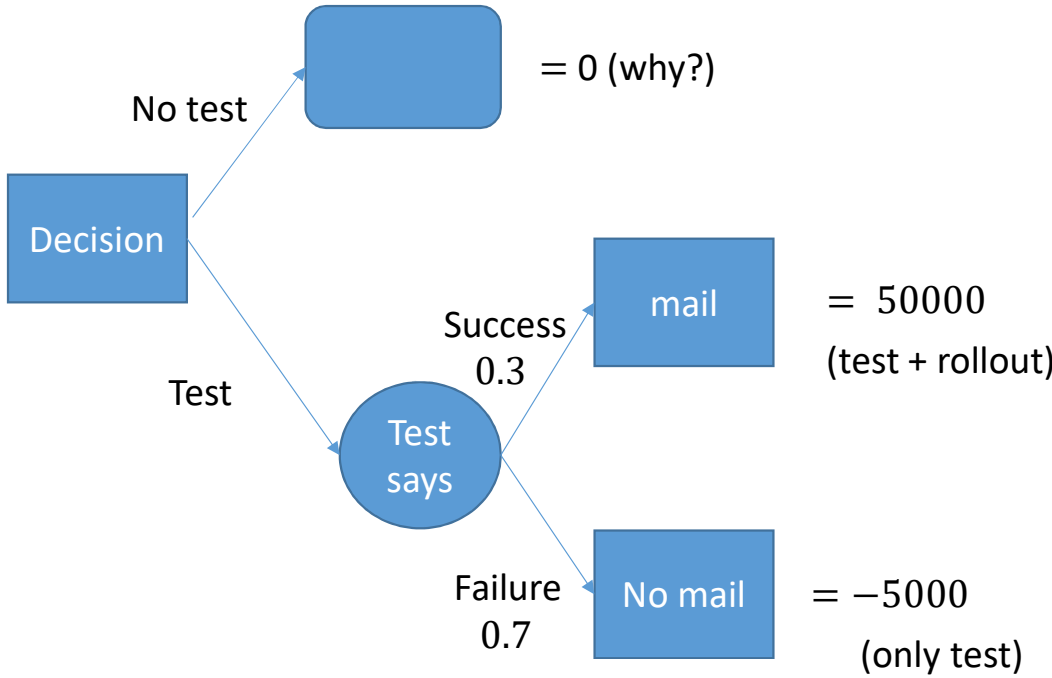
# How big is the option value?

**A1** Assume the test provides perfect information.

**A2** Test predicts

1. Success $(p = .05)$
   $\Rightarrow (m \cdot p - c) = 1.00$
2. Failure $(p = .01)$
   $\Rightarrow (m \cdot p - c) = -1.00$

**A3** Success occurs 30% of the time




$$\text{Value of test} = 0.3 \cdot (50000) + 0.7 \cdot (-5000) - 0$$
$$= \mathbf{11500}$$

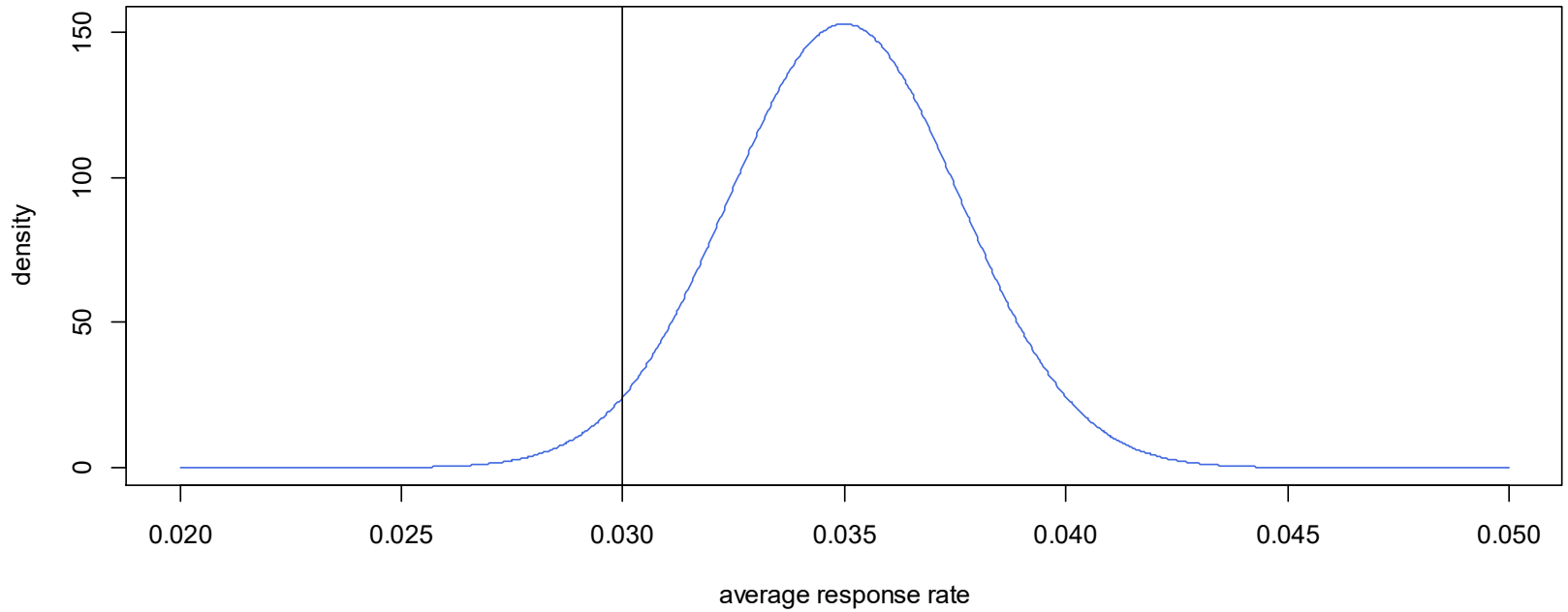
# Uncertainty

The true unobserved population response rate is $p$

- What we observe: sample mean estimate, $\hat{p} = \frac{1}{n}\sum_i x_i$
- Its standard error, $se(p) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{p(1-p)}{n}}$
- Central limit theorem. For large enough sample, distribution of sample mean is approximately normal

$$\hat{p} \sim N(\mathrm{p}, \mathrm{se}(\mathrm{p})^2)$$

# What's the probability we make a mistake?




$$P(\hat{p} < 0.03) = .027$$

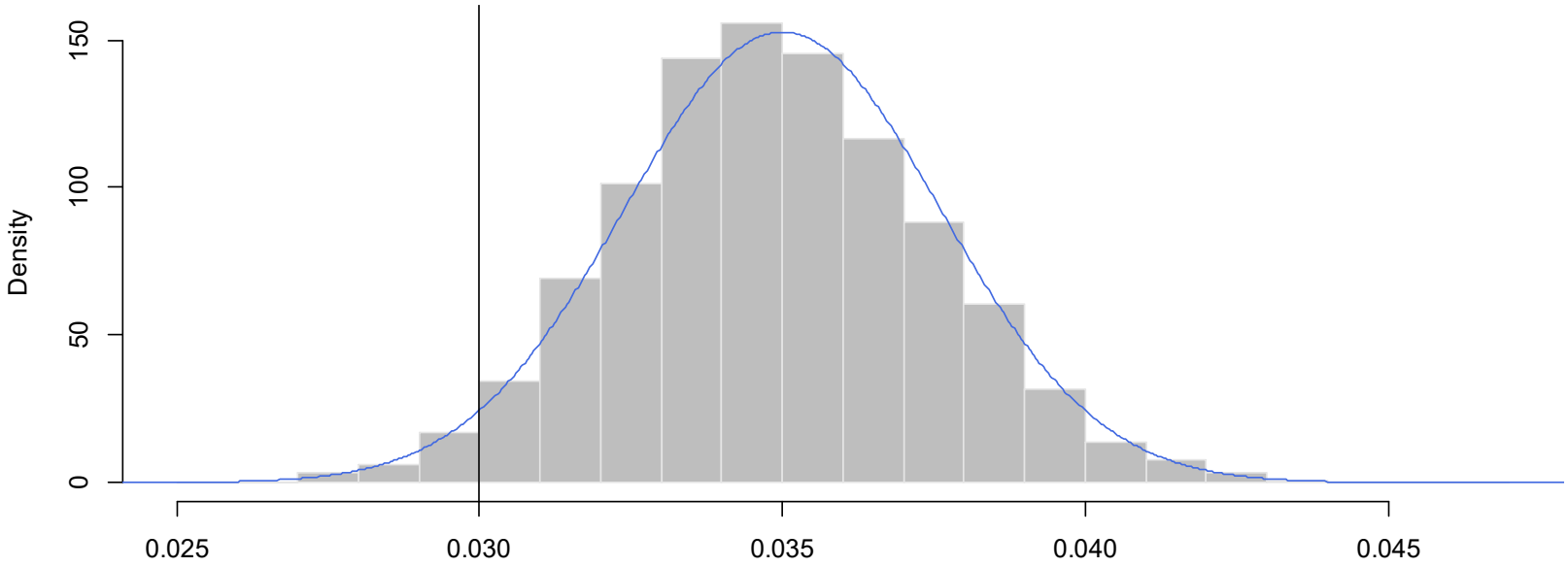
# Bootstrap

Sample *with replacement* from the original sample, using the same sample size

For b = 1 … B bootstrap samples

1. Resample with replacement, $x_1^b, \ldots x_n^b$
2. Calculate estimate using this resample set, $\hat{p}_{\mathrm{b}} = \frac{1}{n}\sum_i x_i^b$

You now have a distribution $\hat{p}_1, \ldots \hat{p}_{\mathrm{B}}$.




$$\frac{\sum_b 1\{\hat{p}_b < .03\}}{B} = 0.0226$$

average response rate

# How big should the test be?

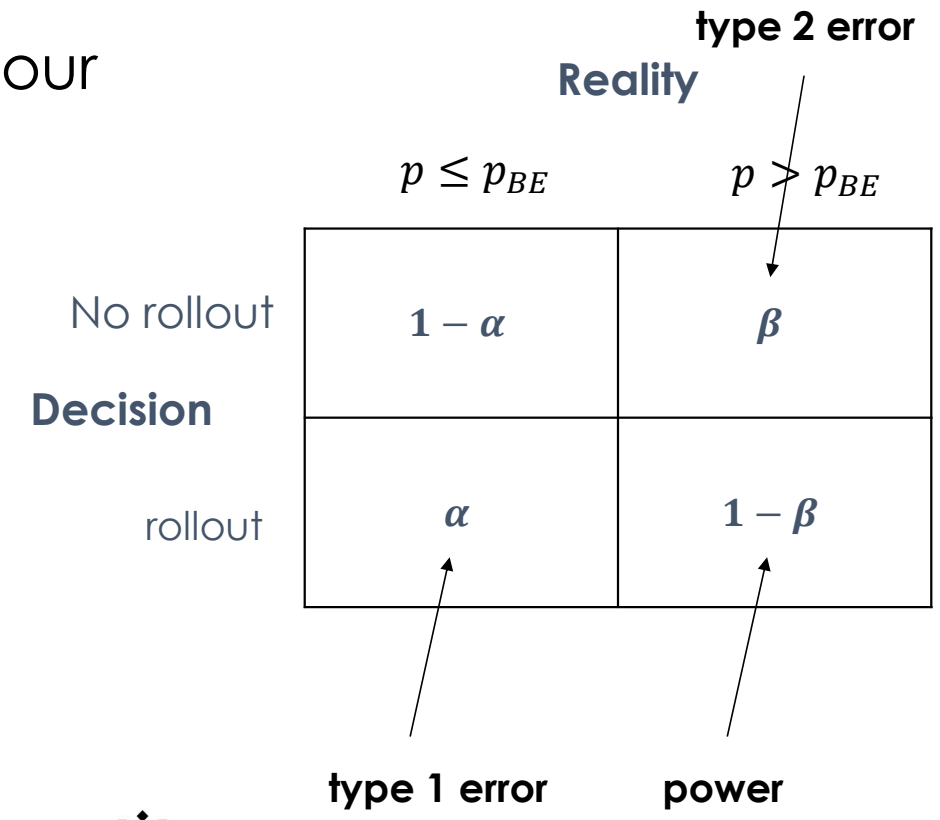
We want to test whether our estimated

$$H_A: p > p_{BE}$$

$$H_0: p \leq p_{BE}$$

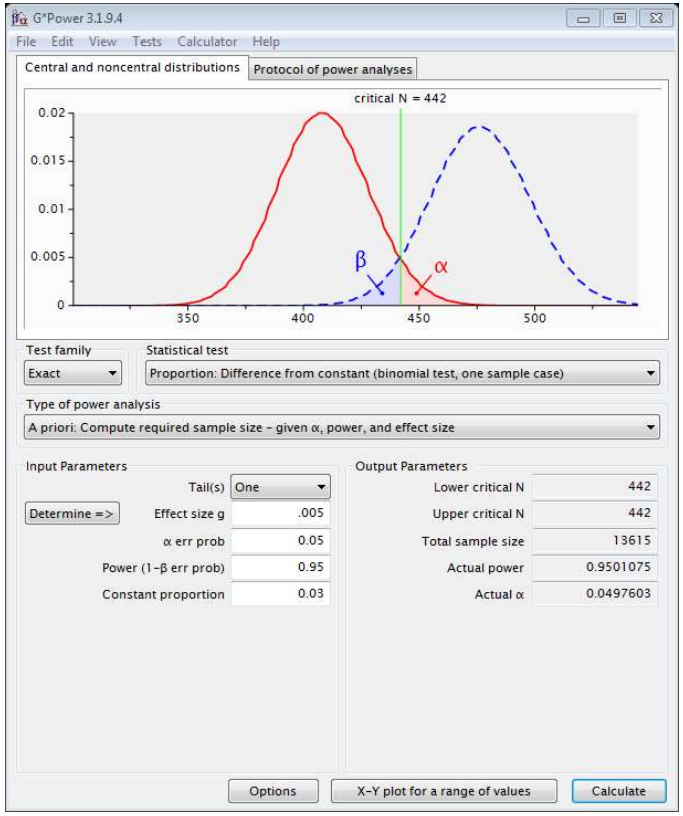
where

$$p_{BE} = \frac{c}{m}$$



# How should you determine the sample size?

- Use excellent (free) software package GPower http://www.gpower.hhu.de/en.html
  - Go to Test family = "Exact"
  - Statistical test = "Proportion: Difference from a constant"
  - Set power $(1 - \beta) = 0.95$
  - Set $\alpha = 0.05$
  - Set constant proportion = $p_{BE}$
  - Set effect size equal to how much over the breakeven your best guess is, $p - p_{BE} = 0.035 - 0.030 = 0.005$



# Making smarter use of the test results

- So far, we've considered an "all or nothing" approach
- What if we used the test to identify profitable groups, and target mailing to them?

# Data for the groups

- Of course we can build a model to select customers on the basis of many variables! Better predictions.

**Most common**

- **Demographics**: gender, ethnicity, age, income, family size, occupation, marital status, education, homeowner or renter, length of residence (typically available for prospects)

- **Transaction data**: past purchases, amounts, dates, discounts, …

**Best but unavailable for prospects**

- **Marketing:** past mailings, content mailings, date, costs

- **(Survey data, e.g. Psychographics)**: attitudes, interests, activities

See BKN chapter 8

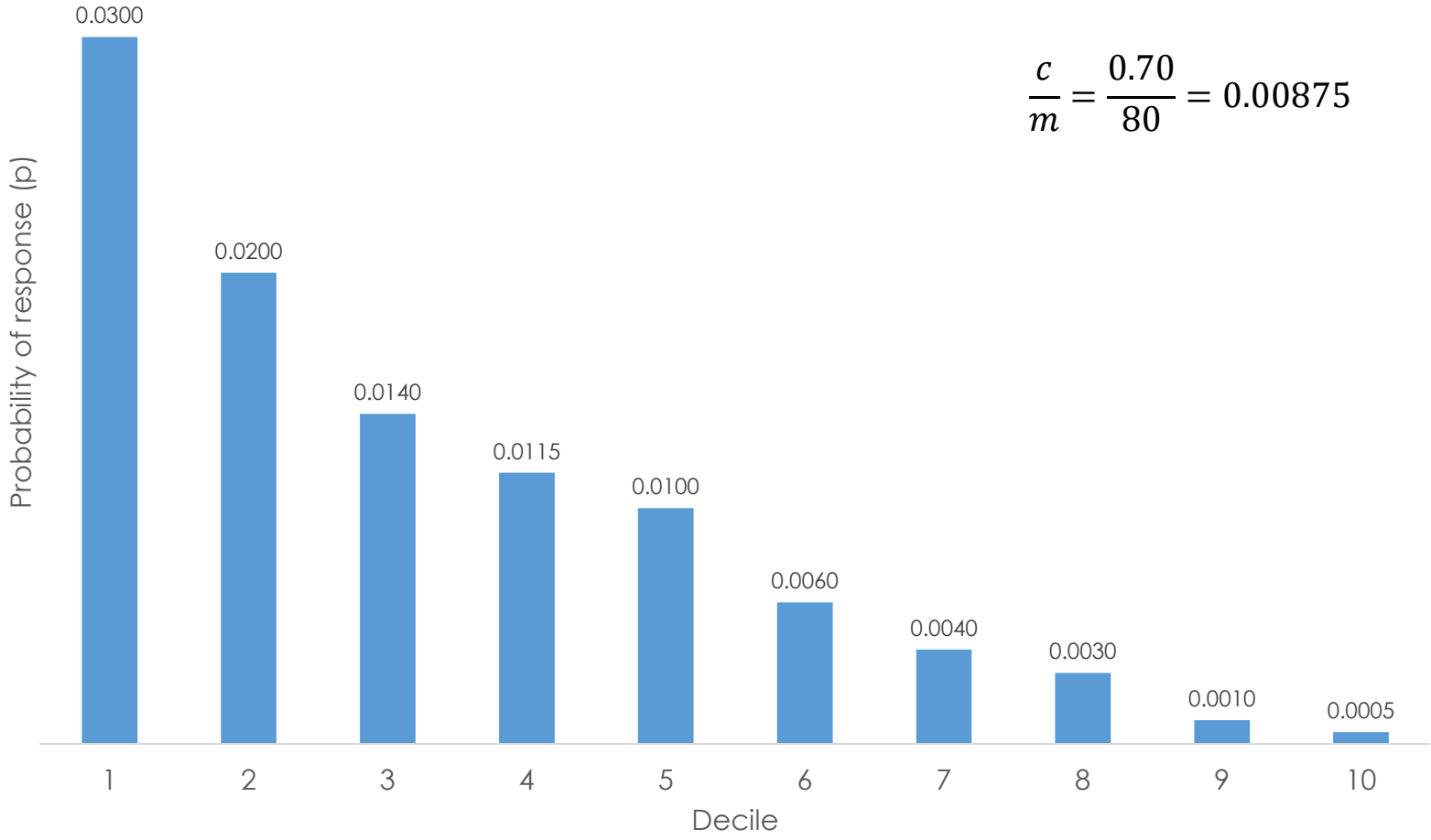
# Example: better targeting with model

- Let's say we use our test data to build a model that predicts response probability
- Profit contribution per response (m) = 80.00
- Each mailing costs 0.70
- Rollout sample size = 1000000

See BKN chapter 8

We sort them into deciles from most likely to least likely

$$\frac{c}{m} = \frac{0.70}{80} = 0.00875$$



# Targeting: select customers based on deciles

| Decile | size | p | (pm - c) | profit | cum profit |
|---|---|---|---|---|---|
| 1 | 100000 | 0.0300 | 1.70 | 170000 | 170000 |
| 2 | 100000 | 0.0200 | 0.90 | 90000 | 260000 |
| 3 | 100000 | 0.0140 | 0.42 | 42000 | 302000 |
| 4 | 100000 | 0.0115 | 0.22 | 22000 | 324000 |
| 5 | 100000 | 0.0100 | 0.10 | 10000 | 334000 |
| 6 | 100000 | 0.0060 | -0.22 | -22000 | 312000 |
| 7 | 100000 | 0.0040 | -0.38 | -38000 | 274000 |
| 8 | 100000 | 0.0030 | -0.46 | -46000 | 228000 |
| 9 | 100000 | 0.0010 | -0.62 | -62000 | 166000 |
| 10 | 100000 | 0.0005 | -0.66 | -66000 | 100000 |

$(.03 * 80 - .70) = 1.70$

# Base case: no targeting

- Rollout to everyone or not at all.

- Response rate to entire list is 0.01

$$E[\text{rollout profit}] = (N - n)(m \cdot p - c)$$
$$= 1000000\left((80)(0.01) - 0.70\right)$$
$$= 100000$$

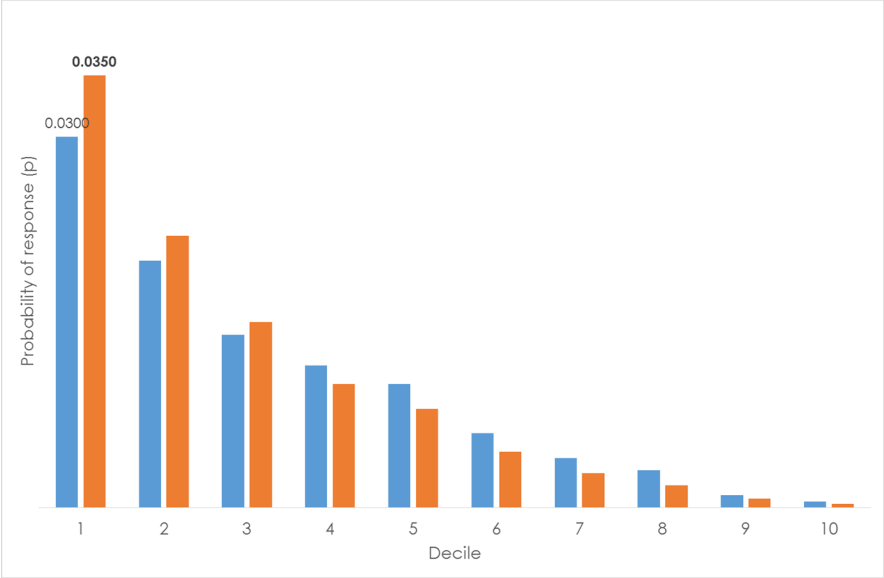
# Summary

- No targeting: mail everyone, profits = 100000
- Targeting: mail top 5 deciles, profits = 334000
  - Cost is 500000 x 0.70 = 350000
  - ROI = profits/cost
    - Mail Top 5 ROI = 334000/350000 = 95%
    - Mail everyone ROI = 100000/700000 = 14%

    For half the cost, three times the profit

Even if the untargeted mailing campaign would be profitable, selecting customers usually is *more* profitable

# Even better predictions?




| Decile | size | p | (pm - c) | profit | cum profit |
|---|---|---|---|---|---|
| 1 | 100000 | 0.0350 | 2.10 | 210000 | 210000 |
| 2 | 100000 | 0.0220 | 1.06 | 106000 | 316000 |
| 3 | 100000 | 0.0150 | 0.50 | 50000 | 366000 |
| 4 | 100000 | 0.0100 | 0.10 | 10000 | 376000 |
| 5 | 100000 | 0.0080 | -0.06 | -6000 | 370000 |
| 6 | 100000 | 0.0045 | -0.34 | -34000 | 336000 |
| 7 | 100000 | 0.0028 | -0.48 | -47600 | 288400 |
| 8 | 100000 | 0.0018 | -0.56 | -55600 | 232800 |
| 9 | 100000 | 0.0007 | -0.64 | -64400 | 168400 |
| 10 | 100000 | 0.0003 | -0.68 | -67600 | 100800 |

ROI = 376000/280000 = 134%

Small improvements in predictions matters!

# Conclusion

- Testing resolves (some, usually not all) uncertainty about the benefit of marketing.
- Testing gives the option to rollout if test results are positive
- Even more value when you use it to better target
- Next class: Recency-Frequency-Monetary (RFM) analysis