Advanced A/B Testing

Small Sample A/B Tests

Elea McDonnell Feit

6/16/2019

When are sample sizes small?

In the age of digital marketing, you'd think we would always have big samples. Not true!

Small sample tests:
- Tests on small sub-populations - Website test on a low-traffic page
- Tests on B2B customers
- Tests where treatments are applied to store locations or geographies - When the treatment is expensive or risky

Wine retailer example

Let's imagine we conducted an activation test with customers who have never purchased, but have been active in the past day.

```
d <- read.csv("test_data.csv")
d <- d[d$past_purch==0 & d$days_since < 1, ]
nrow(d)
```

```
## [1] 187
```

```
aggregate(cbind(open, click, purch) ~ group,
          data=d, FUN=mean)
```

```
##      group       open       click    purch
## 1     ctrl 0.0000000 0.00000000 11.53544
## 2 email_A 0.7866667 0.12000000 43.26760
## 3 email_B 0.7454545 0.07272727 22.63291
```

There are big differences there, but are they statistically significant?

Do emails produce higher purchases?

```
summary(lm(purch ~ email, data=d))
```

```
##
## Call:
## lm(formula = purch ~ email, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34.54 -34.54 -11.54  -2.24 577.74
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.535      8.655   1.333   0.1842
## emailTRUE     23.002     10.380   2.216   0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 65.34 on 185 degrees of freedom
```

Power and precision

We won't always get statistical significance when there are real effects, becasue of noise in the data. **Power** is the likelihood that we will detect a effect when it exists. It is related to of the **precision** of the estimates.

small sample sizes -> low precision and low power

big sample -> use baseline vars to find heterogeneous treatment effects small sample -> use baseline vars to mop up noise

Fix #1: "Regression Correction"

If one of the baseline variables predicts the outcome, then including it in a regression analysis will reduce the error term and increase precision.

$y = a + bx + cz + \varepsilon$

Regression correction for wine retailer

Standard analysis

```
summary(lm(purch ~ email, data=d))$coef
```

```
##                Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 11.53544    8.654864 1.332827 0.18422766
## emailTRUE   23.00210   10.380288 2.215940 0.02791509
```

Regression corrected

```
summary(lm(purch ~ email + visits, data=d))$coef
```

```
##                 Estimate Std. Error    t value    Pr(>|t|)
## (Intercept) -4.372848   13.082853 -0.3342427 0.73857710
## emailTRUE   22.719842   10.336842  2.1979481 0.02920086
## visits       3.309388    2.047807  1.6160642 0.10779370
```

The standard error on emailTRUE may get a bit smaller, but if the baseline variable is unrealated to the outcome, then adding it won't help.

Stratification

Post-stratification

If your customers can be divided into strata that have more homogeneous treatment effects (based on a baseline variable(s)), then you can increase precision/power of your estimate of the overall average treatment effect by computing the estimate separately for each strata and then recombining. (See Berman and Feit, 2018WP for an application in marketing).

Post-stratification is achieved in a regression framework by effects coding the strata indicator (which is a categorical baseline variable) and interacting the treatment indicator with the the baseline indicator.

Post-stratification for the wine retailer

Setup

```
d$strata <- (d$visits < 3)
contrasts(d$strata) <- contr.sum(2)  # effects coding
contrasts(d$strata)
```

```
##       [,1]
## FALSE    1
## TRUE    -1
```

```
summary(lm(purch ~ email*strata, data=d))$coef
```

```
##                      Estimate Std. Error     t value  Pr(>
## (Intercept)        12.395557   13.24082  0.93616243 0.350
## emailTRUE          19.079698   15.17663  1.25717586 0.210
## strata1            -1.140157   13.24082 -0.08610926 0.931
## emailTRUE:strata1   5.994996   15.17663  0.39501485 0.693
```

The wine experiment was completely randomized. The number of subjects who got each treatment is almost perfectly equal. This maximizes the precision.

```
big_d <- read.csv("test_data.csv")
xtabs(~ group, data=big_d)
```

```
## group
##     ctrl email_A email_B
##    41330   41329   41329
```

But we don't have perfect balance between the treatments within group.

```
xtabs(~ group + (visits < 3), data=big_d)
```

```
##            visits < 3
## group       FALSE  TRUE
##    ctrl     37955  3375
##    email_A  37926  3403
##    email_B  37974  3355
```

Stratification (aka blocking)

Balancing the sample in each subgroup will increase precision. This is called (not post-) **stratification** in biostats and **blocking** in engineering. This becomes more important as:

- Sample sizes get very, very small
  - Expensive engineering tests
  - Small populations of patients
  - Tests where stores are the unit of analysis
- The response is very noisy

Matching

Matching

Consider a test where the treatment is a new store display, which we will install in some stores. We usually know lots of things about stores before the experiment.

Matching is when we use the baseline variables that we have on the stores to identify pairs of similar stores.

One of the best ways to match stores is on past sales.

Wait, doesn't this break randomization?

No, we create pairs of matched stores and then randomize within each pair.

Matching in other domains

- **Within-subjects designs** apply both treatments sequentially to each subject
  - Introduces time confounding
  - Reverse the order for some (crossover design)
- **Twin studies** randomly assign twins to different treatments

Paired comparsion signficance test

When you've matched units in advance, you should analyze the test as a paired comparison test.

```
t.test(..., paired=TRUE)
```

Block what you can, randomize what you can not.

atttibuted to George Box, author of *Statistics for Experimenters*

Companies that specialize in small N test design

Mastercard Data & Services Test & Learn

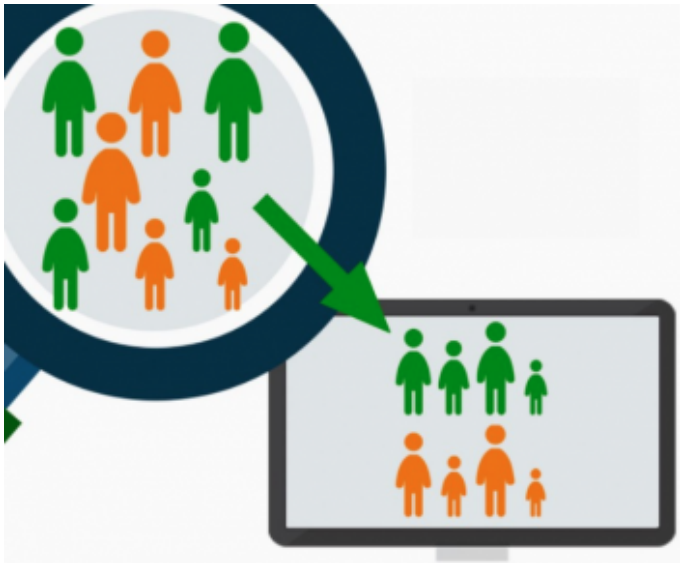
Formerly Applied Predictive Technologies

Google

See Vaver and Koehler 2011

Propensity matching with observational data

Propensity matching attempts to construct an experiment from observational data using match. Matching is constrained by the treatment.



This is still subject to any selection bias related to unobserved variables.

Things you just learned

- Small sample $\rightarrow$ lower power/precision $\rightarrow$ can't find significant differences
- Options for using baseline variables to improve power
  - Add baseline variables as controls ("regression correction")
  - (Post-)stratification ($x \times z$ interactions)
  - Pre-test matching