Date: Thursday, March 28, 2019

... continuing the last example:

$$z^* = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \Big\} \sigma_{12}$$

$H_0: \pi_1 = \pi_2$ vs $H_A: \pi_1 \neq \pi_2$

$H_0: \delta = 0$ vs $H_A: \delta \neq 0$

Reject when $2\,\mathbb{P}(z^* > z_c) \leq \alpha$, where δ = effect size.

↳ Critical value that was determined using α and H_0 true

Under H_0, $z^* \sim N(0,1)$

Under H_A, $\dfrac{p_1 - p_2 - \delta}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$

What is power here?

Power = $\mathbb{P}$(rejecting H_0 | H_A true)

$= 2\,\mathbb{P}(z^* > z_c \mid H_A \text{ true})$

$= 2\,\mathbb{P}\left(\frac{p_1 - p_2}{\sigma_{12}} > z_c \mid H_A \text{ true}\right)$

$= 2\,\mathbb{P}\left(\frac{p_1 - p_2 - \delta}{\sigma_{12}} > z_c - \frac{\delta}{\sigma_{12}} \mid H_A \text{ true}\right)$

$= 2\,\mathbb{P}\left(z > z_c - \frac{\delta}{\sigma_{12}}\right)$ where $z \sim N(0,1)$

Notes: - Power depends on p_1, p_2, n_1, n_2, and δ (effect size)

- When planning an experiment to achieve a desired power, one sets a power for identifying a desired effect size δ.

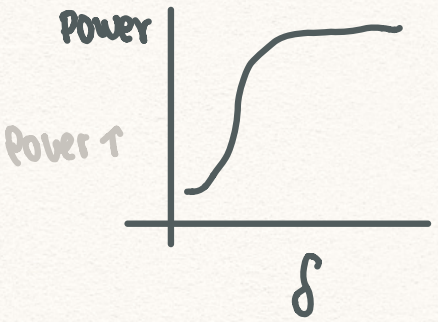
  Ex: Can I detect a difference in CTR between versions A and B of a site of 0.01?

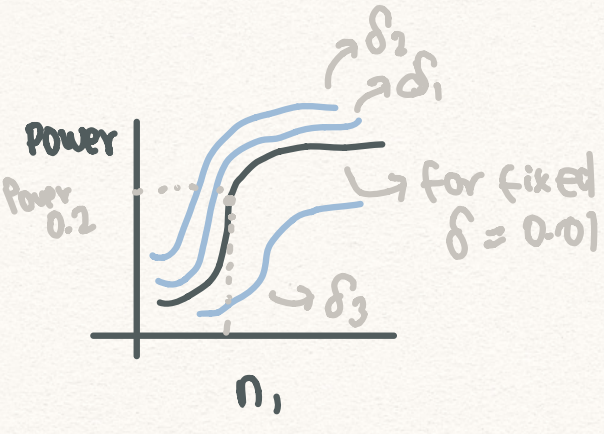
  Strategy — plug in $\delta = 0.01$ and determine n_1 and n_2 needed to obtain a desired power.

♡ Relationships:

1) As $\delta \uparrow$, power $\uparrow$
2) As n_1 or $n_2 \uparrow$, power $\uparrow$

Planning





In general, we state things like "we can detect an effect size of δ with a power 0.8 when $n_1 \geq$ __ and $n_2 \geq$ __".

- p_1 and p_2 also affect our calculations. So in general, we can run a small experiment to estimate p_1 and p_2 or use results of past experiments.

- This was just one example of a z-test. But the same procedure works in general.

  ↳ A rejection region can always be determined by setting $P(\text{Type I error}) \leq \alpha$.

↳ Power required knowing the distribution under H_A, which is not always known, but if not, it is approximated using "fancy footwork" and the CLT.

Types of tests

1) Difference in proportions (π_1 vs π_2)
   - Z-test if n_1, n_2 sufficiently large
     rule of thumb: $n_i \geq 30$
   - Fisher's exact test if n_1, n_2 are small.
2) Difference in means (μ_1 vs μ_2)
   - Z-test if σ_1 and σ_2 are known
   - t-test if σ_1 and σ_2 are unknown

   Note: If σ_1 and σ_2 are different, we need to pool the standard devs to get a test statistic.
3) Differences in variances (σ_1^2 vs σ_2^2)
   - χ^2 test

   Why? $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ = average of squared normal random variables

   $\Rightarrow (n-1)S^2 \sim \chi^2_{n-1}$
4) Goodness of fit and tests of independence
   - χ^2 test or F-test
   - Contingency analysis

   Example coming...
5) Non-parametric tests
   - Testing whether two samples have the same distribution.

   Several approaches:

a) Rank test:

Are the rankings of each sample the same?
ie, if we ordered each sample, would the 3rd largest value be the same in each group? etc.
Specifically, are the quantiles of the samples the same?

b) Empirical Distribution test:

Are the cdfs of the samples the same?

$H_0: F_1(x) = F_2(x)$

$H_A: F_1(x) \neq F_2(x)$

Example:

1) ECDF test via Kolmogorov-Smirnov

Data from group 1: $X_1, \ldots, X_{n_1}$

group 2: $Y_1, \ldots, Y_{n_2}$

Calculate the ECDF for each sample.

Group 1: $F_1(x) = \frac{1}{n_1} \sum_{j=1}^{n_1} 1\{X_j \leq x\}$ = proportion of data $\leq x$

Group 2: $F_2(x) = \frac{1}{n_2} \sum_{j=1}^{n_2} 1\{Y_j \leq x\}$

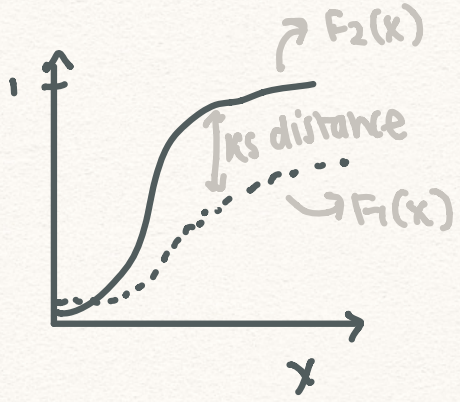
we do this for a grid of x.

$F_1(x)$ is a consistent estimator for $P(X \leq x)$

good asymptotically

KS-stat = $\sup_x \{|F_1(x) - F_2(x)|\}$ = maximum distance between the two distributions

KS-stat is our test statistic which now we can show (through a lot of work) is asymptotically the max of Normal Random variables → we can calculate p-values, etc.



- Why is this useful?

  1) Very easy to implement! ("KS.test" in R)

  2) Does not rely on any parameters!

     ie it provides an answer to whether or not two samples are statistically equivalent

  3) Widely used in econometrics, etc, because there are no underlying modeling assumptions.

Multiple comparisons

- In many cases we want to simultaneously test differences between more than 2 conditions in the same experiment (A/B/n testing).

Example: Testing whether airline prices depend on day of the week, cookies status, time of day.

Factors:
- Day of week
  Levels: M, T, W, ...
- Time of day
  Levels: Morning, Afternoon, Evening, Night
- Cookies
  Levels: Cleared or not

Metric: Cost of United flight 123 from SFO to RDU

Conditions: $\ddot{o}!$ There are a lot!
Let's say there are N of them.

↳ This means that there are $\binom{N}{2}$ different tests we'd like to conduct from our experiment.

Run experiment and get data

X_{ij} = cost under search i from condition j

$i = 1, \ldots, n_j$
$j = 1, \ldots, N$

Tests: $H_{0k}: \mu_j = \mu_{j'}$

$H_{Ak}: \mu_j \neq \mu_{j'}$.

$k = 1, \ldots, m$ = total # of tests

Notation:

m = # hypothesis tested

m_0 = # true null hypotheses

$m - m_0$ = # true alternative hypothesis

V = # false positives (type I errors)

S = # true positives

T = # of false negatives (type II errors)

U = # of true negatives

$R = V + S$ = # rejections.

- How do we assess the type I error associated with a collection of m hypotheses?

- Two primary strategies:
  1) Family-wise error rate (FWER)
  2) False Discovery rate (FDR)

FWER is the probability of making one or more false positives:

$$FWER = \mathbb{P}(V \geq 1)$$

For a single test we set $\mathbb{P}(V=1) \leq \alpha$

we'd like to control FWER to be less or equal to α.

- Goal: Control FWER $\leq \alpha$ for our m tests.

- What if we just do business as usual for each test separately?

$$FWER = \mathbb{P}(V \geq 1) = \mathbb{P}\left(\bigcup_{k=1}^{m} \{\text{rejecting } H_{0k} \mid H_{0k} \text{ true}\}\right)$$

$$\leq \sum_{k=1}^{m} \mathbb{P}(\text{reject } H_{0k} \mid H_{0k} \text{ true})$$

$$= \sum_{k=1}^{m} \alpha = m\alpha$$

no good, especially for large m!

☆ Bonferroni's Method:

For each test, set $\mathbb{P}(\text{Type I error in test } k) \leq \frac{\alpha}{m}$

Bonfer correction!

$\Rightarrow FWER \leq \alpha$, yay!!

- Bonferroni's method states that we set the probability of type I error $\leq \frac{\alpha}{m}$ so that we reject H_{0k} when p-value$_k \leq \frac{\alpha}{m}$. Easy!