# Intro to A/B testing by Udacity

A/B testing isn't that useful testing out new experiences.
Change aversion or Novelty effect.
A/B testing also cannot tell you if you missing something.
Need clear control group and clear metric.

**Click-through-rate** = Number of clicks / Number of page views
**Click-through-probability** = Unique visitors who click at least once / Unique visitors who view the page

## Binomial distribution

- 2 types of outcomes
- Independent events
- Identical distribution, i.e the same P for all
- Mean = P
- Std dev = $\sqrt{\frac{P(1-P)}{N}}$

## How to calculate the Confidence Interval?

$p = \frac{X}{N}$
To use the normal approximation: check $N * \hat{P} > 5$
margin error = $z * \sqrt{\frac{\hat{P}(1-\hat{P})}{N}}$ usually z = 1.96 under 95% C.I

## Hypothesis Testing

### Comparing two samples

- $X_{cont}, X_{exp}, N_{cont}, N_{exp}$
- $\hat{P}_{pool} = \frac{X_{cont}+X_{exp}}{N_{cont}+N_{exp}}$
- $SE_{pool} = \sqrt{\hat{P}_{pool}(1-\hat{P}_{pool}) * (\frac{1}{N_{cont}} + \frac{1}{N_{exp}})}$
- $\hat{d} = \hat{P}_{exp} - \hat{P}_{cont}$
- $H_0 : \hat{d} = 0$ and $\hat{d} \sim N(0, SE_{pool})$
- If $\hat{d} > 1.96 * SE_{\text{pool}}$ or $\hat{d} < -1.96 * SE_{pool}$, reject the $H_0$

### An example

$X_{cont} = 974, X_{exp} = 1242, N_{cont} = 10072, N_{exp} = 9886$, the minimal detectable difference required is 2%. Under 95% C.I, would you launch the new feature?
**Ans:** We can compute the 95% C.I of $\hat{d}$ as follows:
$\hat{P}_{pool} = (974 + 1242)/(10072 + 9886) = 0.111$
$SE_{pool} = \sqrt{0.111 * (1 - 0.111) * (\frac{1}{10072} + \frac{1}{9886})} = 0.00445$
$\hat{d} = \frac{1242}{9886} - \frac{974}{10072} = 0.0289$
Marginal error = $m = SE_{pool} * 1.96 = 0.0087$
Therefore the 95% C.I is given by $(\hat{d} - m, \hat{d} + m) = (0.0202, 0.0376)$. Because the lower bond is greater than the required minimal detectable difference 2% so we will launch it!
**When to launch and when not?**
If the C.I upper bound < $d_{min}$, not launching it.
Else if the C.I lower bound > $d_{min}$, launch it.
Else additional tests are required.

### Statistics

- $\alpha = P(\text{reject null}|\text{null is true})$
- $\beta = P(\text{fail to reject}|\text{null is false})$
- Sensitivity = Statistical power = $1 - \beta$, often set to 80%
- Small sample: $\alpha$ is low but $\beta$ is high
  Imagine you increase the sample size, the std error would be smaller --> narrower distribution.
- Online sample size calculator

## Principles

Our recommendation is that there should be internal reviews of all proposed studies by experts regarding the questions:

- Are participants facing more than minimal risk?
- Do participants understand what data is being gathered?
- Is that data identifiable? (e.g. timestamp data may be considered as sensitive information)
- How is the data handled?

## Metrics

Start with high level concepts for metrics. e.g. Funnel analysis.

- **Define** metrics by utilizing external data or internal data (e.g. user research and user surveys)
- **Filter** out spam and fraud data (traffic). e.g. a big press coverage increases the website traffic or the change only impacts subset of your traffic. How to decide the filtering criteria technically? Computing *baseline* value for your metrics! Also we can compute the metrics segment by some factors such as language or platform. When you apply the filtering, check if you are moving traffic disproportionately from one fo these places or not. It's a good thing or a bad thing. It might align on the business **intuitions** (for say all spam come from some country) or you might bias your results further. It's important to build intuitions so we know if the changes are expected or not.
- Summary metrics such as it's a number (e.g. load time of a video; what the position of the first click on the page is), so you can choose a whole set of metrics such as mean, median, 25th percentile, 75th percentile. How to choose between them?
  - Establish a few characteristics for your metric, **Sensitivity and Robustness**, in the sense that you want your metric sensitive enough in order to detect a change. Also it's robust enough to not picking up the changes you don't care about. For example, mean could be misleading due to outliers. Median is not very good in the loading time example in the sense that the change probably affect 20% of people (it's huge) but unable to move the median much.
  - Run **A/A test** to test sensitivity and robustness.
  - Also we could do a *retrospective analysis* and plot a histogram of the summary metric (e.g. loading time, checking its distribution). If it's more like a normal, using mean or median. If it becomes more one sided, or lopsided, maybe consider 25th, 75th or 90th percentiles.
- Sum/counts
- Rate/probabilities
- Ratio
- Calculate the variability
  - probability -> binomial distribution
  - mean -> normal -> sample mean statistics has the variance $\frac{\sigma^2}{N}$ and given the the **Central Limit Theorem**, the sample mean will approach a normal distribution for a large $N$. This is true for a sample of independent random variables from **any** population distribution, as long as the population has a finite standard deviation $\sigma$.
  - median/percentiles -> depends the underlying data distribution
  - count/difference -> normal
  - rates -> poisson -> $\text{Var}(X) = E(X) = \hat{X}$
  - ratios -> depends on the underlying data distribution
- Calculate the variability by non-parametric methods when the distribution is very wired.
  - Run a lot of A/A tests to estimate the empirical variability and C.I
  - Bootstrapping

### An example

- Group A, Group B. We can compute the difference between two groups. If we assume the difference follows a normal distribution, then 95% C.I is given by $\text{mean(diff)} + -(1.96) \times \text{std(diff)}$
- On the other hand, if we don't believe it follows the normal distribution, we can just sort the difference by value and manually pick the 95 C.I threshold from the histogram (e.g. 40 samples then 95% C.I would be throwing out the biggest one and the smallest one)

## Design the experiments

- Define "subject"
  - user id: stable, unchanging, personally identifiable
  - anonymous id such as cookie: changes when you switch browser or device, users can clear cookie
  - event id: no consistent experience
  - device id
  - IP address
- Define "population"
  - Normally A/B testing means **"inter-user experiments"**, i.e. you get different people on group A and and on group B side
  - Sometime you can do **"intra-user experiments"** where you show A and B to the same user at different time. But choosing the timing is tricky.
  - **Cohort** analysis: cohort basically means that people who entered the experiment at the same time. Matching up people in two groups so we have roughly the same parameters in two user groups. Cohort v.s. population: use Cohort when you are looking for user stability, having a *learning effect* or you want to measure something like increased usage of the site or examining user retention or churn
- Size
- Duration: proportional to the exposure, i.e. the traffic assigned to the experiment group and control group. Sometimes prefer to run an experiment on small percentage of people for long period of time

## Analyzing the results

- Sanity checks: checking invariant metric
- Single metric v.s. Multiple metrics: more things you test, the more likely you are to see significant differences just by chance. To address it:
  - Bonferoni correction, very conservative because usually the metrics are correlated with each other. $p_i < \frac{\alpha}{\#\text{ of hypothesis tests}}$
  - FDR correction controls false positive rate, recommended, widely used in genome wide studies where often thousands of hypothesis tests are conducted simultaneously. FDR = False Discovery Rate = Expected (# false predictions/ # total rejections against null). The FDR is the rate that statistic test called significant are truly null. An FDR of 5% means that, among all tests called significant, 5% of these are truly null.
  - FWER: controls probability that *any* metric shows a false positive
- How to draw conclusions?
  - Have a composite metric such as RPM (revenue per thousand queries)
  - Come up with a good Overall Evaluation Criteria (OEC) that balanced long term and short time benefits.
  - What if I have some statistically significant metrics while having other non-significant ones? --> Do you understand the change? Experiences from other experiments might help.
  - What if the metric is statistically significant in some segments but not in other segments? --> Do you understand why?
  - Start with the small traffic and if we see significant results then gradually ramp-up to the whole population. However, the effect may be flatten out as you ramp yo the change due to some reasons like the holiday effect, seasonality etc. One way to capture these event-driven impact is to use "holdback". The idea is that we launch the change to everyone except for the a small holdback, a set of users, that don't get the chance.
- **Simpson's paradox**: statistically significance appears in several different groups of data but disappears or reverses when these groups are combined.