Overview of A/B Testing

Customer funnel:

Homepage visits → Exploring the site → Create account → Reach some sort of completion

Audacity example: change the "Start Now" button from orange to pink

Metric:

- Total number of courses completed? No. Using this metric would simply take too much time to be practical
- Number of clicks? No. What happens if more total users view the page in one version of the experiment?
- Click-through-rate: number of clicks / numbers of page views. No. It is used when you want to measure the usability of the site. It tells how often the users find that button.
- Click-through-probability: unique visitors who click / unique visitors to page. It is used when you want to measure the impact of the site. It tells how often the users went to the second level page on the site.

Hypothesis: changing the button will increase the click-through-probability of the button.

Click-through-probability： binomial distribution

- 2 types of outcomes (success, failure)
- Independent events
- Identical distribution (P same for all)

If we know the click-through-probability should follow binomial distribution, we can use the formula we have for sample standard error for the binomial to estimate how variable we expect our overall probability of a click to be. What that means is that for a 95% confidence interval, if we theoretically repeated the experiment over and over again, we would expect the interval we construct around our sample mean covers the true value in the population of 95% of the time.

The center of the confidence interval:

Estimated success probability: P(hat)= X/N.

A good rule of thumb to tell whether a binomial distribution is normal if the sample is large enough:

1. N*P(hat) > 5 (this's the more stringent condition for small probability problems); 2. N* (1-P(hat)) > 5

The width of the confidence interval if we can use the normal approximation:

The margin of error: m = z score of the confidence level * standard error

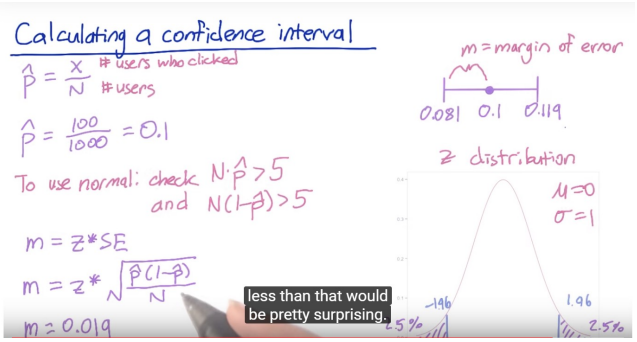
$= z * \sqrt{\frac{Phat(1-Phat)}{N}}$ (* standard error for binomial distribution)

Marginal error, which is the amount of random variation we expect in our sample and the width of the confidence interval, is a function of both the proportion of successes and the sample size. This means we need to consider the proportion of success when deciding how many samples to collect.

When the success probability p is farther from 0.5, SE and CI will be smaller, the distribution is tighter.

Similarly, if the number of samples is larger, the SE and CI will also be smaller.

For a normal distribution with a mean of 0 and a standard deviation of 1, with 95% confidence, the true value would be with 1.96 and -1.96 of the estimates we observed. Since we're doing a 2-tailed test, each tail will contain 2.5% of distribution. So, 1.96 is the z-score for 97.25%.



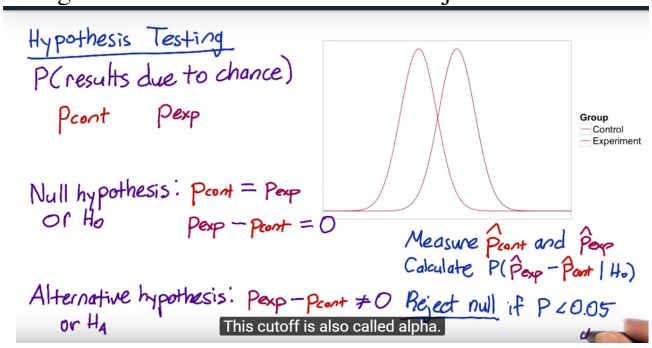
Hypothesis testing/ Statistical inference: a quantitative way to establish how likely it is that the results occurred by chance.
We'll need to compare the proportion of clicks estimated on the control side and the experiment side and measure whether the difference we observed could have occurred by chance, or if it would be extremely unlikely to have occurred if the two sides were actually the same, which means that the difference is statistically significant.
In order to calculate the probability that your results are due to chance, you need to have a hypothesis about what the results would be if your experiment had no effect.
Null hypothesis (baseline):

> There's no difference in CTP between our control and our experiment, which means we would expect the 2 groups to have equivalent distributions, so they would be right on top of each other. Pcont – Pexp = 0

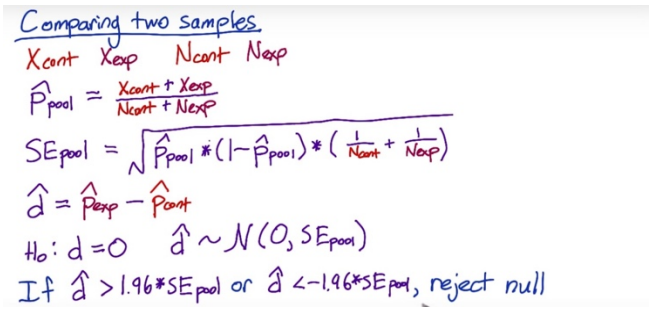
Alternative hypothesis:

> The experiment does have an effect. We expect Pcont – Pexp != 0

We can estimate Pcont and Pexp from the data we collected and then calculate the difference between these (Pcont - Pexp) and compute the probability that this difference would have arisen by chance if the null were true (P(Pcont – Pexp | H0)).
Then we want to reject the null, and conclude that our experiment has an effect if this probability is small enough. We can choose the cutoff of rejection at 0.05.



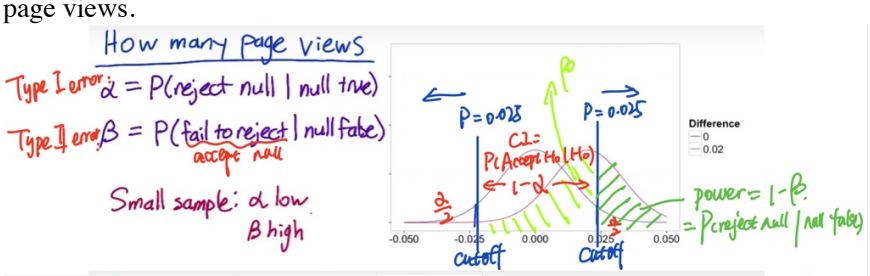
Because we have 2 samples, we'll need to choose a standard error that gives us a good comparison of both. The simplest thing we can do is calculate a pooled standard error.
X: the number of users who click in each group
N: total number of users in each group
P(hat) is the pooled probability, which is the total probability of a click across groups

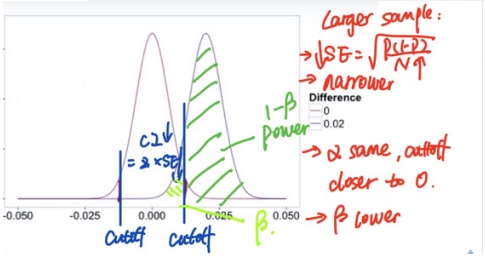


We would expect our estimation of difference between groups to be distributed normally, with a mean of 0 and a standard deviation of the pooled standard error.

Then, we should decide what change in the probability is practically significant from a business perspective. What size change matters to us?
What you really want to observe is repeatability. And you want to make sure when you set up your experiment that you get that guarantee that these results are repeatable, so it's statistically significant. But you also want to make sure that if you can see change in your experiment that you're interested in from a business standpoint, so it's practically significant and also statistically significant. You need to size your experiment appropriately, such that the statistical significance bar is lower than the practical significance bar.

Design the experiment:
Statistical power is the main question we have to decide: given that we have control over how many page views go into our control and experiment, we have to decide how many page views we need in order to get a statistically significant result. If we see something interesting, we want to make sure that we have enough power to conclude with probability that the interesting result is statistically significant.
Power has an inverse trade-off with size. The smaller the change that you want to detect, or the increased confidence interval that you want to have in the result, means that you have to run a larger experiment, so more page views.

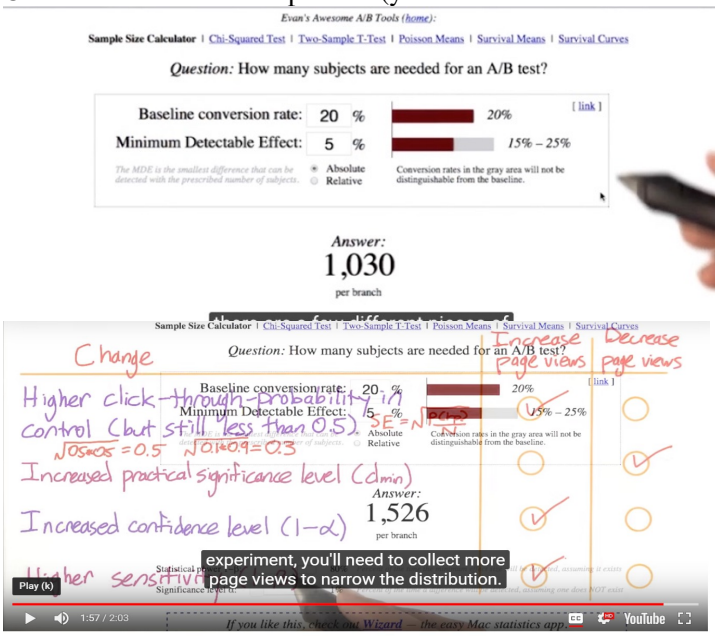




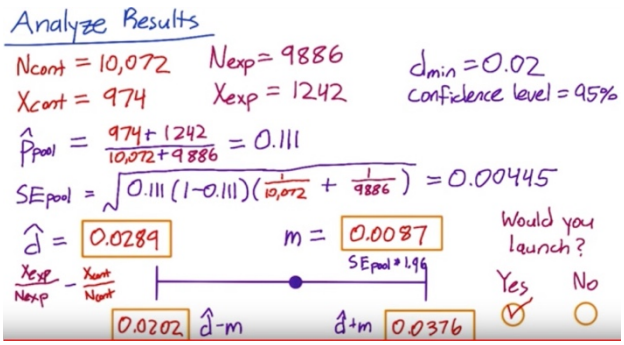
Blue curve is the distribution of results would look like if you collected 1000 samples and there was no true difference between the groups because of a 0 mean. Your probability of falsely concluding there was a difference is alpha: 0.05 (Type 1 error).
If you increased your sample size, the SE would decrease, so the distribution will look narrower. To keep alpha the same, the cutoffs for rejecting will be closer to 0.
Red curve refers to the case where there was difference\. The probability of failing to reject the null when the null is false is beta (Type 2 error), which is pretty high with small sample size.
By collecting a small sample, alpha is low, that is, you're unlikely to launch a bad experiment. But beta is high, that is, you're likely to fail to launch an experiment that actually did have a difference you care about.
For normal distribution, as your true change gets larger, beta will go down, that is, a lower chance of error.

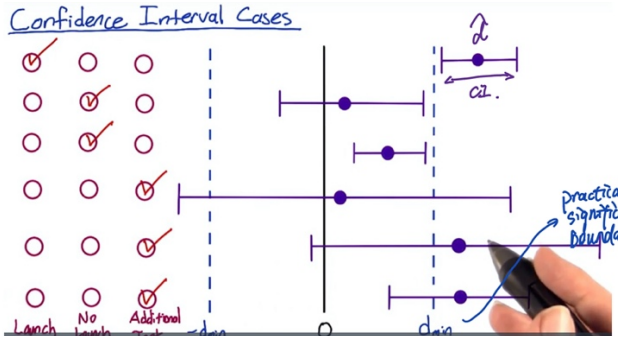
1 – beta = = P (reject H0 | Ha) = sensitivity/power. In general, you want your experiment to have a higher level of sensitivity at the practical significance boundary, which is always set as 80%
For larger samples, alpha doesn't change. In the case where there's a true difference, you're much less likely to fall within the range of failing to reject the null, that is, you're more likely to reject the null and conclude there was a difference. Beta has gone down and power increased.

Online calculator for sample size (you can also use built-in library):



Baseline: the estimated CTP before making the change
Minimum dateable effect: practical significance level
#1. Probability level gets closer to 0.5 → sqrt(p*(1-p)) increases → standard error increases → means that I'd also need to increase the number of page views(N) in order to reduce the SE back to its original level.
#2. Larger changes are easier to detect than smaller changes, so do not need that many samples.
#3. Increase confidence interval – you want to be more certain that a change has occurred before you reject the null. More conservative – more samples needed (or you can reject the null less often, but power will decrease)
#4. Increase the power – need to collect more samples to narrow the distribution.





Case #2: Neutral case: CI includes 0. There's not a practically significant change.
Case #3: Upper bound of CI is less than practical significance, and CI does not include 0. The change is statistically significant, but the magnitude is not large enough for you to care about.
Case #4 - #6: We need to run additional tests with greater power to draw conclusions if time allows. If don't have time, we should communicate to the decision-makers when they're going to have to make a judgement, and take a risk, because the data is uncertain. They're going to have to use other factors, like strategic business issues, or other factors besides the data.