Designing an A\B Test

Udacity Data Analyst Nanodegree - Project 7

Diego Menin – November 2015

UDACITY

Project Overview

This is the seventh project on the Udacity Data Analyst NanoDegree. On this project, I will consider an actual experiment that was run by Udacity where I'll flesh the experiment idea out into a fully defined design, analyse the results, and propose a high-level follow-on experiment.

This project is connected to the A/B Testing course and built based on the submission template found here. The submission consist on a series of questions (that you will see in the beginning of each section in blue) that I'll be answering though this document. It is not this project's objective to create a step-by-step documentation or explanation of an A\B test. If something seems "out of context" please refer to the template for further instructions.

This project doesn’t require any explicit coding so all the calculations were done using google spreadsheets. I tried to present\document most of the intermediate steps.

This document is divided on the following section:

Contents

Project Overview 2

Experiment Overview: Free Trial Screener 3

Experiment Design 4

Metric Choice 4

Measuring Standard Deviation 6

Sizing 7

Number of Samples vs. Power 7

Duration vs. Exposure 8

Experiment Analysis 9

Sanity Checks 10

Result Analysis 10

Recommendation 13

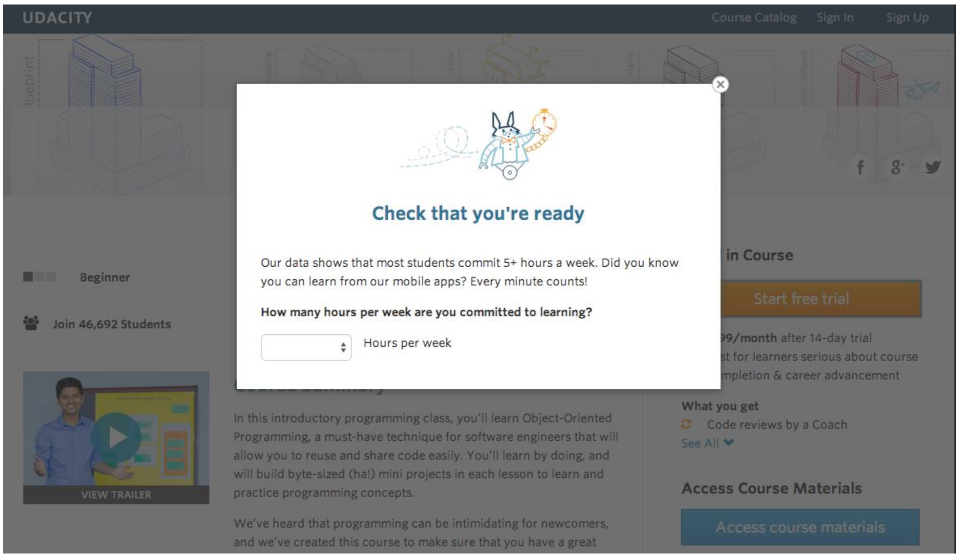
Follow-Up Experiment 14

Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead.

This screenshot shows what the experiment looks like.



Check that you're ready

Our data shows that most students commit 5+ hours a week. Did you know you can learn from our mobile apps? Every minute counts!

How many hours per week are you committed to learning?

Hours per week

In this introductory programming class, you'll learn Object-Oriented Programming, a must-have technique for software engineers that will allow you to reuse and share code easily. You'll learn by doing, and will build byte-sized (ha!) mini projects in each lesson to learn and practice programming concepts.

We've heard that programming can be intimidating for newcomers, and we've created this course to make sure that you have a great

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit of diversion is a cookie, although if the student enrols in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here.

The metrics chosen for this experiment are:

> **Invariant Metrics:** Number of Cookies, Number of Clicks
> **Evaluation Metrics:** Gross Conversion, Retention, Net Conversion

Metrics Explanation:

Number of cookies:
Represents the number of unique cookies to view the course overview page.
Chosen as invariant metric because the number of cookies are independent from the experiments since the visits happen before the user sees the experiment.

Number of clicks:
Represents the number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger).
Chosen as an invariant metric for the same reason as the number of cookies. It happens before the experiment, so it will be the same for control and experiment group (equal probability of clicking the Start Free Trial button by both groups)

Click-through-probability:
Represents the number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.
Not Chosen as an invariant metric because it is the ratio between two variable who are already selected as Invariant Metrics.

Gross conversion:
Represents the number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button.
Chosen as an evaluation metric because it directly depends on the effect of the experiment

Retention:
Represents the number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout.
Chosen as an evaluation metric because it directly depends on the effect of the experiment and indicates financial gain resulted from the change

Net conversion:
Represents the number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button.

Chosen as an evaluation metric because it directly depends on the effect of the experiment and indicates financial gain resulted from the change

Number of user-ids:
Represents the number of users who enrol in the free trial.
It is a bad invariant metric because user-ids aren't tracked unless the student enrols, so this is actually equivalent to number of enrolments, which could be different between control and experiment.
I would be usable as evaluation metric because it would track the second part of the hypothesis, namely that we won't reduce the number of students to continue past the free trial but it was not chosen as such because it is not normalized.

Expectations:

The experiment consists on presenting a pop up to the user right after clicking the "Start Free Trial" button, asking if he\she can devote more than 5 hours per week to it.

By doing that, I expect to reduce the number of users that abandon the course right after starting the free trial due to lack of time, bringing down Gross Conversion.

At the same time, I expect Net conversion to have at least a statistically significant increase, because it takes into consideration the number of user who make at least one payment thus directly affecting our revenue.

So, to launch the experiment, I will require Gross conversion to have a practically significant decrease, and Net conversion to have a statistically significant increase.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

Bellow I'll list the standard deviation of each of the evaluation metrics. For each of them, I indicate whether the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different. The standard deviation will be calculated using the formula:

$$\sigma = (p\ (1 - p)/n)^{1/2}$$

That's the standard deviation of a binomial distribution and refer to 40000 unique cookies view the page per day. From those, 3200 will click on the "Start free trial" button (0.08% of them). That's the click-through-probability on "Start free trial" button;

On the current course format, from all the 3200 clicks, 660 will enrol on the course, which gives a probability of enrolling of 0.20625 %.

Again, on the current course format, the probability of payment is 0.53%, which results on the probability of payment of 0.1093125 %

Unique cookies to view page per day:	40000
Unique cookies to click "Start free trial" per day:	3200
Click-through-probability on "Start free trial":	0.08
Enrolments per day:	660

Gross Conversion	Probability of enrolling, given click:	0.20625
Retention	Probability of payment, given enroll:	0.53
Net Conversion	Probability of payment, given click	0.1093125

Given a Sample size of 500:

	Numerator	Denominator	Ratio	Square Root (Standard Deviation)
	$P\ (1 - p)$	n		
Gross Conversion	0.163710938	400	0.000409	**0.02023**
Retention	0.2491	82.5	0.003019	**0.05495**
Net Conversion	0.097363277	400	0.000243	**0.01560**

In choosing whether to use an analytical or empirical estimate we should check if the denominators, or the unit of analyses, are cookies as it always is for our unit of diversion. In this project, "cookies" are chosen as our unit of diversion.

Since both our evaluation metrics, Gross conversion and net conversion, have the number of cookies as their denominator, we could then presume that the analytical estimate will probably be accurate and we would not need to resort to the empirical one in that case.

We can then proceed using an analytical estimate of the variance.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of page views you will need to power you experiment appropriately.

	Baseline conversion rate	dmin	Samples Needed	Page Views	times 2 (Experiment and Control groups)
Gross Conversion	20.625%	1.00%	25,835	322,938	645,875
Retention	53.000%	1.00%	39,115	2,370,606	4,741,212
Net Conversion	10.931%	0.75%	27,413	342,663	685,325

Considerations:

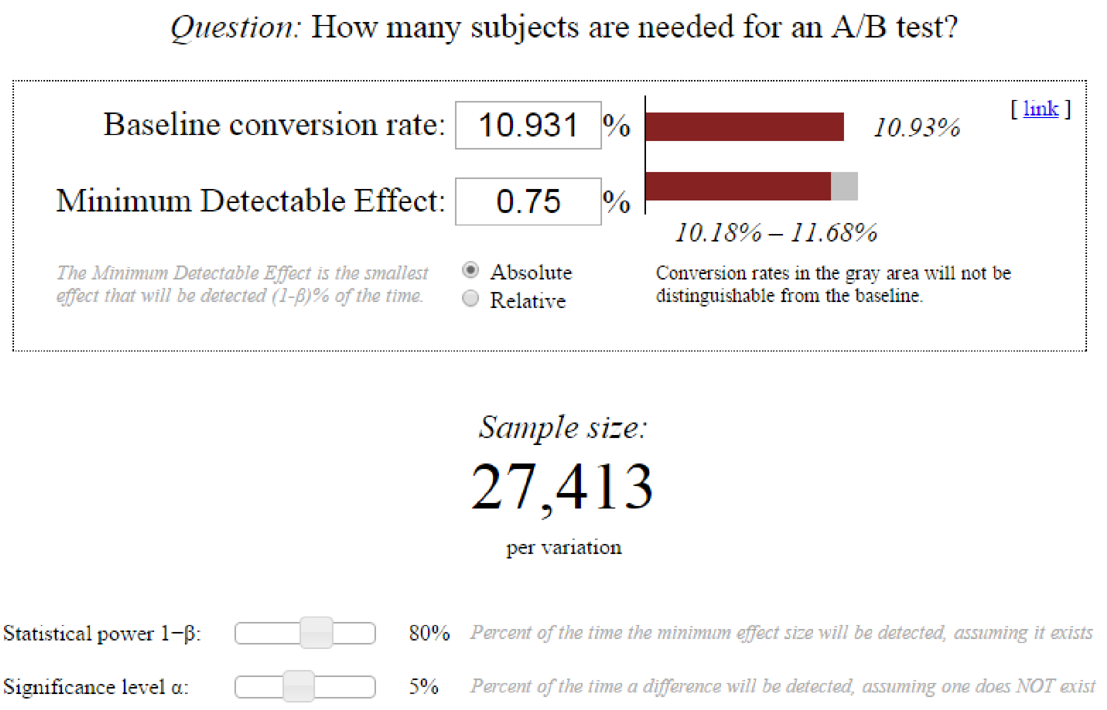
- Base Line Conversion rate: click-through-probability before making the change;
- dmin: Minimal detectable Effect: practical significance level;
- The evaluation metrics I selected to proceed with are Gross conversion and Net conversion.
- Since to evaluate retention would need over 8M page views, I've decide to use Gross Conversion and Net Conversion so 685 324 page views will be necessary;

I am not applying the Bonferroni correction because, in order to launch the experiment, I would need both metrics to be relevant, rather than one out of two.

The case where all metrics need to be significant in order to launch is not the same as the case where any metrics can be significant. In fact it is the exact opposite. For the former the risk of a type two error increases as the number of metrics increase, for the latter the risk of a type I error increases.

If I were to launch the experiment when ANY metric would be significant (so we would launch if just one proves significant) then I would have to use Bonferroni because, out of 2 metrics, the risk that just one would be significant by pure chance (Type I error) would be very high so in order to reduce that risk we apply Bonferroni.

The "samples needed" where calculated using this online calculator:



Question: How many subjects are needed for an A/B test?

Baseline conversion rate: 10.931 %

Minimum Detectable Effect: 0.75 %

The Minimum Detectable Effect is the smallest effect that will be detected (1-β)% of the time.

● Absolute
○ Relative

Conversion rates in the gray area will not be distinguishable from the baseline.

Sample size:

27,413

per variation

Statistical power 1−β: 80% *Percent of the time the minimum effect size will be detected, assuming it exists*

Significance level α: 5% *Percent of the time a difference will be detected, assuming one does NOT exist*

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment.

I would divert 100% of the traffic to the experiment, which will result on the experiment taking 18 days, which is a reasonable time. I chose that because I don’t consider the experiment to be risky given that:

- It does not affect existing customers (either paying or non-paying);
- Is very simple, affecting only one step of the workflow;

Experiment Analysis

The data for the analysis can see below. (Original spreadsheet here). Note that on the spreadsheet there are two sheets - one for the experiment group, and one for the control group. I merged them on the same table for easy of use. The meaning of each column is:

- Page views: Number of unique cookies to view the course overview page that day.
- Clicks: Number of unique cookies to click the course overview page that day.
- Enrolments: Number of user-ids to enroll in the free trial that day.
- Payments: Number of user-ids who enrolled on that day to remain enrolled for 14 days and thus make a payment. (Note that the date for this column is the start date, that is, the date of enrolment, rather than the date of the payment. The payment happened 14 days later. Because of this, the enrolments and payments are tracked for 14 fewer days than the other columns.)

Control						Experiment				
Date	Pageviews	Clicks	Enrollments	Payments		Date	Pageviews	Clicks	Enrollments	Payments
Sat, Oct 11	7723	687	134	70		Sat, Oct 11	7716	686	105	34
Sun, Oct 12	9102	779	147	70		Sun, Oct 12	9288	785	116	91
Mon, Oct 13	10511	909	167	95		Mon, Oct 13	10480	884	145	79
Tue, Oct 14	9871	836	156	105		Tue, Oct 14	9867	827	138	92
Wed, Oct 15	10014	837	163	64		Wed, Oct 15	9793	832	140	94
Thu, Oct 16	9670	823	138	82		Thu, Oct 16	9500	788	129	61
Fri, Oct 17	9008	748	146	76		Fri, Oct 17	9088	780	127	44
Sat, Oct 18	7434	632	110	70		Sat, Oct 18	7664	652	94	62
Sun, Oct 19	8459	691	131	60		Sun, Oct 19	8434	697	120	77
Mon, Oct 20	10667	861	165	97		Mon, Oct 20	10496	860	153	98
Tue, Oct 21	10660	867	196	105		Tue, Oct 21	10551	864	143	71
Wed, Oct 22	9947	838	162	92		Wed, Oct 22	9737	801	128	70
Thu, Oct 23	8324	665	127	56		Thu, Oct 23	8176	642	122	68
Fri, Oct 24	9434	673	220	122		Fri, Oct 24	9402	697	194	94
Sat, Oct 25	8687	691	176	128		Sat, Oct 25	8669	669	127	81
Sun, Oct 26	8896	708	161	104		Sun, Oct 26	8881	693	153	101
Mon, Oct 27	9535	759	233	124		Mon, Oct 27	9655	771	213	119
Tue, Oct 28	9363	736	154	91		Tue, Oct 28	9396	736	162	120
Wed, Oct 29	9327	739	196	86		Wed, Oct 29	9262	727	201	96
Thu, Oct 30	9345	734	167	75		Thu, Oct 30	9308	728	207	67
Fri, Oct 31	8890	706	174	101		Fri, Oct 31	8715	722	182	123
Sat, Nov 1	8460	681	156	93		Sat, Nov 1	8448	695	142	100
Sun, Nov 2	8836	693	206	67		Sun, Nov 2	8836	724	182	103
Mon, Nov 3	9437	788				Mon, Nov 3	9359	789		
Tue, Nov 4	9420	781				Tue, Nov 4	9427	743		
Wed, Nov 5	9570	805				Wed, Nov 5	9633	808		
Thu, Nov 6	9921	830				Thu, Nov 6	9842	831		
Fri, Nov 7	9424	781				Fri, Nov 7	9272	767		
Sat, Nov 8	9010	756				Sat, Nov 8	8969	760		
Sun, Nov 9	9656	825				Sun, Nov 9	9697	850		
Mon, Nov 10	10419	874				Mon, Nov 10	10445	851		
Tue, Nov 11	9880	830				Tue, Nov 11	9931	831		
Wed, Nov 12	10134	801				Wed, Nov 12	10042	802		
Thu, Nov 13	9717	814				Thu, Nov 13	9721	829		
Fri, Nov 14	9192	735				Fri, Nov 14	9304	770		
Sat, Nov 15	8630	743				Sat, Nov 15	8668	724		
Sun, Nov 16	8970	722				Sun, Nov 16	8988	710		
Total	345,543	28,378	3,785	2,033			344,660	28,325	3,423	1,945

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data.

First, I started by checking if the numbers of assignments on the control group is the same as the numbers on the experiment group:

Page views:
Control: 345543
Experiment: 344660

The Control\Experiment allocation can be thought of a random event with exactly two outcomes (probability of assignment = 0.5) – binomial distribution. So we'll need to calculate the binomial confidence interval. The Total sample size is 690203, which is enough to assume a normal distribution. (The binomial distribution assumes a normal distribution when n is large). I will go step by step for the "Page View" calculation only:

1. Compute the standard deviation of a binomial distribution with probability of 0.5 of success:

   SD = 0.0006

2. Calculate the margin of error with a 95% confidence interval:

   m = SD * 1.96
   m = 0.001180

3. Calculate the confidence Interval (5% of the time the observed values should fall into this range):

   CI = [0.5 –m, 0.5 +m]
   CI = [0.4988, 0.5012]

4. Check if the observed values is within the interval:

   Values on the control group:
   1 – (344,660 / 690,203) = 0.5006

Here's the summary:

	Total	p	1-p	/ n	SD	Margin of Error @95	Lower Bound	Upper Bound	Value on the Control Group
Pageviews	690,203	0.5	0.5	0.00000036221	0.0006	0.001180	0.4988	0.5012	0.5006
Clicks	56,703	0.5	0.5	0.00000440894	0.0021	0.004116	0.4959	0.5041	0.5005

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant.

We can only understand whether or not a user have gone through the full enrolment process until the 2nd of November, that’s why the total clicks on the formulas will only include that interval

	Clicks	Enrollments	Payments
CONTROL	17293	3785	2033
EXPERIMENT	17260	3423	1945
Total	34553	7208	3978

For Gross Conversion, we calculate the “p – experiment” by dividing number of enrolments by number of clicks on the experiment totals and the “p – control” by doing the same calculation on the control Group. The Net Conversion follows the same logic but it uses the number of payments on the numerator:

	p - experiment	p - control	d
Gross Convertion	0.19832	0.21887	-0.02055
Net Convertion	0.11269	0.11756	-0.00487

Then we calculate the pooled probability by using the same formula on the total values and the standard error is calculated by the formula:

$$\sigma = (p\ (1 - p)/n)^{½}$$

n in this case being:

$$(1/n\text{-}control \quad + 1/n\text{-}experiment)$$

because the sample size on the control is different than the sample size on the experiment.

Next we calculate the margin of error by multiplying the SE by the z-score for 95% confidence (1.96) and apply the margin of error to the “d” to get the Lower and Upper CI boundaries:

Pooled Probability	SE	Margin of Error	Lower CI	Upper CI
0.2086	0.0044	0.0086	-0.0291	-0.0120
0.1151	0.0034	0.0067	-0.0116	0.0019

Considering that the practical significance (d_min) for gross conversion is +/- 0.01 and for Net Conversion is 0.0075:

Gross conversion: Is statistically significant because the CI range doesn’t contain zero and practically significant because the CI doesn’t contain the d_min value.

Net conversion: Is not statistically significant because the CI range contains zero and is not practically significant because the CI contains the d_min value.

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant

To do a sign test using the day-by-day data, we need to know how many successes we've for each measure and the total number of trials. Success can be considered when the metric was bigger on the experiment than on the control. Looking at the table above we can see that there were 4 successes for Gross Conversion and 10 successes for Net conversion. The number of trials is 23 in both cases.

Using an online calculator like this, we can plug in the values and get the two-tailed p value, which represents the chance of observing either X or fewer successes, or Y-X or more successes, in Y trials by chance.

	Sucesses	Trials	Probability	two-tailed p-value
Gross Conversion	4	23	0.5	0.0026
Net Conversion	10	23	0.5	0.6776

Since the p-value is smaller than the alpha of 0.5, the sign test agrees with the hypotheses test, that this result is unlikely to have unlikely to have come about by chance.

Recommendation

Make a recommendation and briefly describe your reasoning.

Two hypothesis tests were performed on this experiment, one on “Gross Conversion” and the other on Net Conversion”.

Gross Conversion went down by at least the practical significance boundary. This is a good outcome because we want to decrease the cost (whether that is monetary, or the cost of unsatisfied students) of enrolments that aren't likely to stick it out.

As for the Net Conversion there was no statistically significant change, but the confidence interval does include the negative of the practical significance boundary. That is, it's possible that this number went down by an amount that would matter to the business. Therefore, there is a risk that if we go ahead with the experiment we may end up with a decrease in revenue.

The question is: Is this an acceptable risk in order to launch? And it is mainly a business decision, but I’d advise against it. The practical gain of this experiment (cost decrease on the enrolments) won’t be worth it if we end up losing money by having less students making at least one payment.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

In order to increase the number of “paying students”, I believe Udacity could follow a different approach when it comes to the “start free trial” process. Currently, when you select “start free trial”, you are prompted to input your credit card (CC) information, even though the course is free for the first 14 days.

I believe that many students decide not to do it simply because they don’t want to inform their CC information on such an early stage, especially if they are not sure they’ll purchase the product (and there are many reason for that, they may think they’ll get improper billing or that they even may forget to cancel and be charged for a service they are not using).

The way I see it should work is: the user should be able to take the 14-day trial without informing their CC information (all other user information would still be required to register) and only after the trial ended, be asked if he\she wants to continue, then being prompted for the CC information.

So my hypothesis would be:

- Not requesting user’s credit card information when starting the free trial will increase the Net Conversion of the course.

The “Invariant Metric” would be “number of cookies” because the experiment happens after registration and it is conducted only with user_id, for the experiment group, after providing identification, CC details would not be asked for control group they would.

The Evaluation metric would be “Net Conversion”, the ratio between users that make one payment and the number of users who started the free trial.

The unit of diversion would be “user id” since we’ll still be asking for user information on the “Start Free Trial”.

References:

https://en.wikipedia.org/wiki/Binomial_distribution

https://en.wikipedia.org/wiki/Bernoulli_distribution

http://www.evanmiller.org/ab-testing/sample-size.html

http://graphpad.com/quickcalcs/binomial1.cfm

http://stats.stackexchange.com/questions/154542/what-exactly-is-multiple-testing-bonferroni-correction-can-i-use-it-with-a-si

http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf