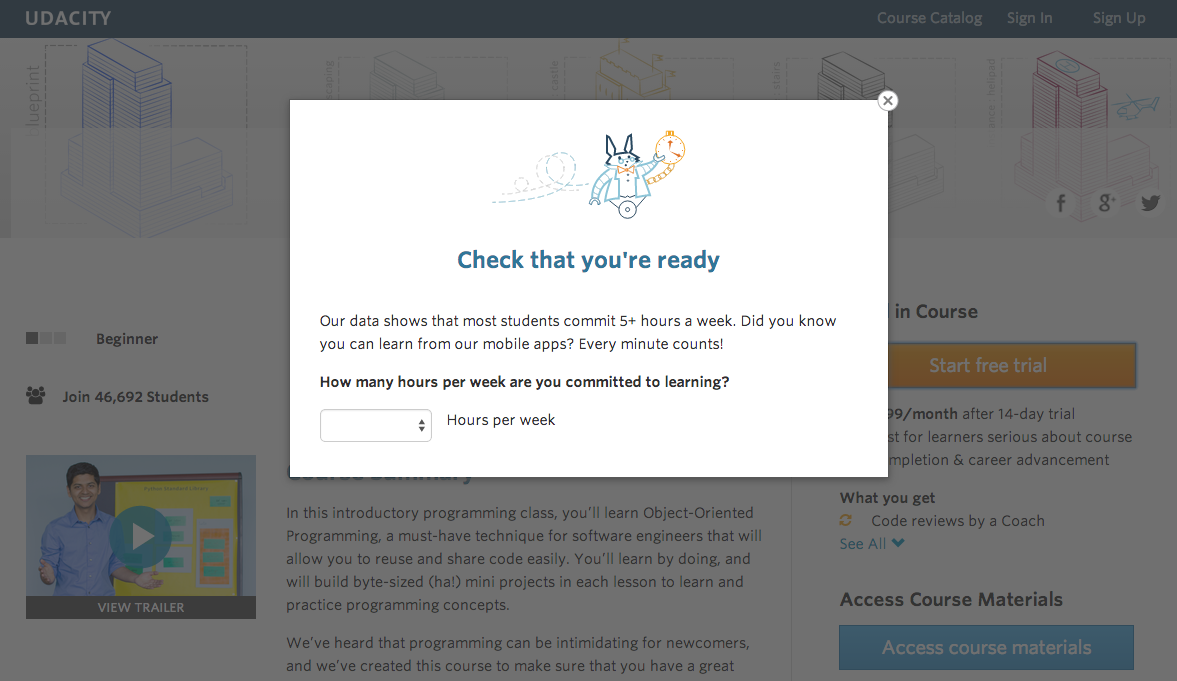
# Design an A/B Test

Experiment Design

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 the experiment:



The primary aim of Udacity is to improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

**Null Hypothesis** : The null hypothesis is that this approach might not make a significant change and might not be effective in reducing the early Udacity course cancellation.

**Alternative Hypothesis** : The alternative hypothesis is that this might reduce 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.

**Metric Choice**

Any place "unique cookies" are mentioned, the uniqueness is determined by day. (That is, the same cookie visiting on different days would be counted twice.) User­ids are automatically unique since the site does not allow the same user­id to enroll twice.

● Number of cookies: That is, number of unique cookies to view the course overview page. (dmin=3000)

● Number of user­ids: That is, number of users who enroll in the free trial. (dmin=50)

● Number of clicks: T hat is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (dmin=240)

● Click­through­probability: That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. (dmin=0.01) ● Gross conversion: T hat is, number of user­ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button. (dmin= 0.01)

● Retention: That is, number of user­ids to remain enrolled past the 14­day boundary (and thus make at least one payment) divided by number of user­ids to complete checkout. (dmin=0.01) ● Net conversion: That is, number of user­ids to remain enrolled past the 14­day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button. (dmin= 0.0075)

#### Invariant metrics (expected to be unchanged in the control and experimental groups)

#### number of cookies (cannot be affected by the experiment: users made a decision to visit the page before they were asked the question);

#### number of clicks (cannot be affected by the experiment: users clicked the button before they were asked the question);

#### click-through probability (cannot be affected by the experiment: it equals to the number of clicks divided by the number of cookies).

#### Evaluation metrics (expected to be different in the control and experimental groups):

#### gross conversion (can be affected by the experiment / can decrease: users could make a decision to enroll in the free trial in the experimental group less than in the control group because they did not plan to learn 5+ hours per week);

#### retention (can be affected by the experiment / can increase: enrolled users could be disappointed in the learning process less and make more payments in the experimental group than in the control group because they paid attention to studying 5+ hours per week);

#### net conversion (can be affected by the experiment / can decrease: users could enroll in the free trial less in the experimental group than in the control group, thus could decrease the number of people who paid).

#### Unused Metrics

#### Number of user-ids: The number of users who enroll in the free trial. User-ids are tracked only after enrolling in the free trial and equal distribution between the control and experimental branches would not be expected. User-id count could be used to evaluate how many enrollments stayed beyond the 14 day free trial boundary, but since it isn't normalized, I have elected not to use it.

#### The goals of the experiment in the practical meaning:

#### the number of payments should not be decreased;

#### the number of students who were disappointed and had not paid because they could not study enough time should be reduced.

#### The goals of the experiment in terms of our metrics:

#### the gross conversion should significantly decrease;

#### the retention should significantly increase;

#### the net conversion should not decrease.

**Measuring Standard Deviation**

This list contains rough estimates of the baseline values for these metrics (again, these numbers have been changed from Udacity's true numbers).

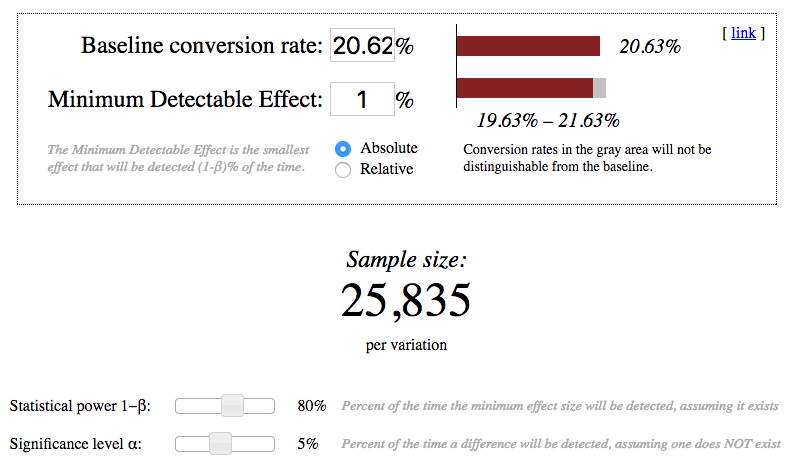
* Unique cookies to view page per day: 40000
* Unique cookies to click "Start free trial" per day: 3200
* Enrollments per day: 660
* Click-through-probability on "Start free trial": 0.08
* Probability of enrolling, given click: 0.20625
* Probability of payment, given enroll: 0.53
* Probability of payment, given click 0.1093125
* Number of cookies = 5000
* Number of clicks on "Start free trial" = 5000 × 0.08 = 400
* Number of enrollments = 5000 × 0.08 × 0.20625 = 82.5
* "SD Gross conversion = ", math.sqrt(0.20625 \* (1 - 0.20625) / 400)
* "SD Retention = ", math.sqrt(0.53 \* (1 - 0.53) / 82.5)
* "SD Net conversion = ", math.sqrt(0.1093125 \* (1 - 0.1093125) / 400)
* SD Gross conversion = 0.020230604137
* SD Retention = 0.0549490121785
* SD Net conversion = 0.0156015445825

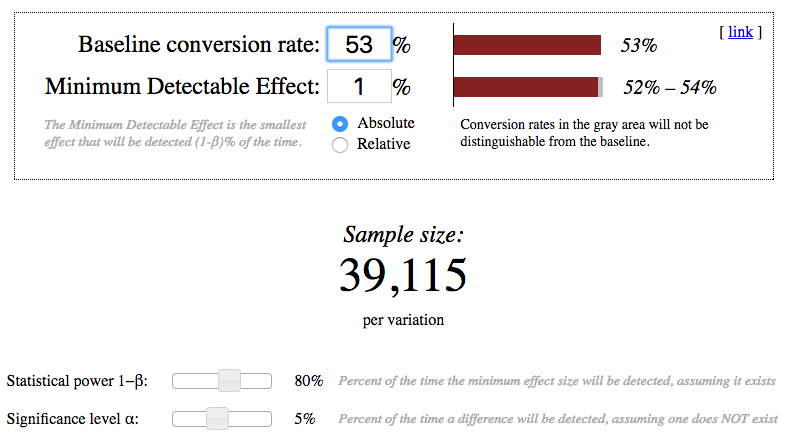
**Sizing**

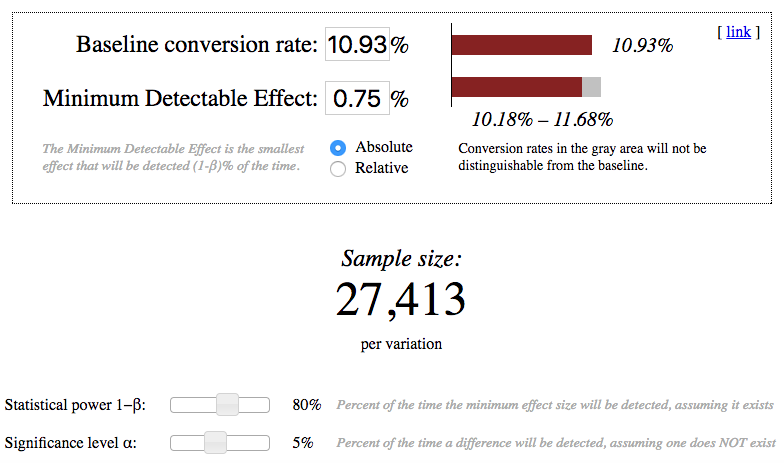
**Number of Samples vs. Power**

❓ Using the analytic estimates of variance, how many pageviews total (across both groups) would you need to collect to adequately power the experiment? Use an alpha of 0.05 and a beta of 0.2. Make sure you have enough power for each metric.

❗ I have used the [online calculator](http://www.evanmiller.org/ab-testing/sample-size.html) for calculating the sample sizes and have chosen the largest.







Gross conversion: 2 × 25835 × 40000 ÷ 3200 = 645875.0

Retention: 2 × 39115 × 40000 ÷ 660 = 4741212.12121

Net conversion: 2 × 27413 × 40000 ÷ 3200 = 685325.0

Pageviews required is maximum of pageviews required for Gross Conversion, Retention, Net Conversion. Therefore, the required pageviews is 47,41,212.

**Duration vs. Exposure**

If we divert 100% of traffic, given 40,000 page views per day, the experiment would take ~ 119 days. If we eliminate retention, we are left with Gross Conversion and Net Conversion. This reduces the number of required pageviews to 685,325, and an ~ 18 day experiment with 100% diversion and ~ 35 days given 50% diversion.

For the experiment with the gross conversion and the net conversion, we can use the period 17 days with the 100% traffic level. This interval is much better, but it gives us the result too quickly. The behavior of people in the field of education is quite difficult to analyze and trends in this area could be rarely detected in a short time period for 2-3 weeks. To slightly increase the time interval, we will set the percentage of used traffic at 60 (fraction = 0.6) and it gives us the number: 29 days.

Experiment Analysis

**Sanity Checks**

For the counts metrics, we assumed that 50% of the experiment traffic goes to the experiment group and 50% goes to the control group. If we call these two groups success and failure, then the model can be a Bernoulli distribution. So I would check whether the current counts of these two groups can come from a population with 0.5 change of success or failure.

p = 0.5

**print** "Number of cookies:"

SE\_cookies = math.sqrt(p \* p / (345543 + 344660))

**print** "Standard error SE = ", SE\_cookies

ME\_cookies = SE\_cookies \* 1.96

**print** "Margin of error ME = ", ME\_cookies

(LB\_cookies, UB\_cookies) = (p - ME\_cookies, p + ME\_cookies)

**print** "Confidential interval CI = ", (LB\_cookies, UB\_cookies)

**print** u'p**\u0302**', "=", 345543.0 / (345543 + 344660), u'**\u2208**', (LB\_cookies, UB\_cookies), u'**\u2713**'

**print**

**print** 'Number of clicks on “Start free trial":'

SE\_clicks = math.sqrt(0.5 \* 0.5 / (28378 + 28325))

**print** "Standard error SE = ", SE\_clicks

ME\_clicks = SE\_clicks \* 1.96

**print** "Margin of error ME = ", ME\_clicks

(LB\_clicks, UB\_clicks) = (p - ME\_clicks, p + ME\_clicks)

**print** "Confidential interval CI = ", (LB\_clicks, UB\_clicks)

**print** u'p**\u0302**', "=", 28378.0 / (28378 + 28325), u'**\u2208**', (LB\_clicks, UB\_clicks), u'**\u2713**'

**print**

**print** 'Click-through-probability on "Start free trial":'

p\_pool = 1.0 \* (28378 + 28325) / (345543 + 344660)

**print** "Pooled probability p\_pool = ", p\_pool

SE\_pool = math.sqrt(p\_pool \* (1 - p\_pool) \* (1.0 / 345543 + 1.0 / 344660))

**print** "Standard error SE = ", SE\_pool

ME\_pool = SE\_pool \* 1.96

**print** "Margin of error ME = ", ME\_pool

d\_hat = 28325.0 / 344660 - 28378.0 / 345543

**print** "Difference", u'd**\u0302**', "= ", d\_hat

(LB\_pool, UB\_pool) = (0 - ME\_pool, 0 + ME\_pool)

**print** "Confidential interval CI = ", (LB\_pool, UB\_pool)

**print** u'd**\u0302**', u'**\u2208**', (LB\_pool, UB\_pool), u'**\u2713**'

Number of cookies:

Standard error SE = 0.000601840740294

Margin of error ME = 0.00117960785098

Confidential interval CI = (0.49882039214902313, 0.5011796078509769)

p̂ = 0.500639666881 ∈ (0.49882039214902313, 0.5011796078509769) ✓

Number of clicks on “Start free trial":

Standard error SE = 0.0020997470797

Margin of error ME = 0.00411550427621

Confidential interval CI = (0.49588449572378945, 0.5041155042762105)

p̂ = 0.500467347407 ∈ (0.49588449572378945, 0.5041155042762105) ✓

Click-through-probability on "Start free trial":

Pooled probability p\_pool = 0.0821540908979

Standard error SE = 0.000661060815639

Margin of error ME = 0.00129567919865

Difference d̂ = 5.66270915869e-05

Confidential interval CI = (-0.0012956791986518956, 0.0012956791986518956)

d̂ ∈ (-0.0012956791986518956, 0.0012956791986518956) ✓

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.

(These should be the answers from the "Sanity Checks" quiz.)

#### 🔵   Number of cookies

Lower bound = 0.4988; Upper bound = 0.5012; Observed = 0.5006; Passes = ✅

#### 🔵   Number of clicks on “Start free trial"

Lower bound = 0.4959; Upper bound = 0.5041; Observed = 0.5005; Passes = ✅

#### 🔵   Click-through-probability on "Start free trial" (Difference between the control and experimental groups)

Lower bound = -0.0013; Upper bound = 0.0013; Observed = 0.0001; Passes = ✅

**Result Analysis**

**Effect Size Tests**

##### *Check for Practical and Statistical Significance*[*¶*](https://olgabelitskaya.github.io/P7_Design_an_A_B_Test_Overview.html#Check-for-Practical-and-Statistical-Significance)

❓ Next, for your evaluation metrics, calculate a confidence interval for the difference between the experiment and control groups, and check whether each metric is statistically and/or practically significance. A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)

In [18]:

**print** "Control group:"

**print** "Clicks = ", control\_data2['Clicks'].sum(), " ", \

"Enrollments = ", control\_data2['Enrollments'].sum(), " ", \

"Payments = ", control\_data2['Payments'].sum()

**print** "Experimental group:"

**print** "Clicks = ", experiment\_data2['Clicks'].sum(), " ", \

"Enrollments = ", experiment\_data2['Enrollments'].sum(), " ", \

"Payments = ", experiment\_data2['Payments'].sum()

Control group:

Clicks = 17293 Enrollments = 3785.0 Payments = 2033.0

Experimental group:

Clicks = 17260 Enrollments = 3423.0 Payments = 1945.0

In [19]:

*# Supporting calculation for checking evaluation metrics*

**print** "Gross conversion"

p\_pool = 1.0 \* (3785.0 + 3423.0) / (17293 + 17260)

**print** "Pooled probability p\_pool = ", p\_pool

SE\_pool = math.sqrt(p\_pool \* (1 - p\_pool) \* (1.0 / 17293 + 1.0 / 17260))

**print** "Standard error SE = ", SE\_pool

ME\_pool = SE\_pool \* 1.96

**print** "Margin of error ME = ", ME\_pool

d = 3423.0 / 17260 - 3785.0 / 17293

**print** "Difference d", "= ", d

(LB\_pool, UB\_pool) = (d - ME\_pool, d + ME\_pool)

**print** "Confidential interval CI = ", (LB\_pool, UB\_pool)

**print** (-0.01, 0, 0.01), u'**\u2209**', (LB\_pool, UB\_pool)

**print** "Statistical significance", u'**\u2713**', " Practical significance ", u'**\u2713**'

**print**

**print** "Net conversion"

p\_pool = 1.0 \* (2033.0 + 1945.0) / (17293 + 17260)

**print** "Pooled probability p\_pool = ", p\_pool

SE\_pool = math.sqrt(p\_pool \* (1 - p\_pool) \* (1.0 / 17293 + 1.0 / 17260))

**print** "Standard error SE = ", SE\_pool

ME\_pool = SE\_pool \* 1.96

**print** "Margin of error ME = ", ME\_pool

d = 1945.0 / 17260 - 2033.0 / 17293

**print** "Difference d = ", d

(LB\_pool, UB\_pool) = (d - ME\_pool, d + ME\_pool)

**print** "Confidential interval CI = ", (LB\_pool, UB\_pool)

**print** 0, u'**\u2208**', (LB\_pool, UB\_pool), "; dmin = -0.0075", u'**\u2208**', (LB\_pool, UB\_pool)

**print** "Statistical significance", u'**\u2718**', " Practical significance ", u'**\u2718**'

Gross conversion

Pooled probability p\_pool = 0.208607067404

Standard error SE = 0.00437167538523

Margin of error ME = 0.00856848375504

Difference d = -0.0205548745804

Confidential interval CI = (-0.0291233583354044, -0.01198639082531873)

(-0.01, 0, 0.01) ∉ (-0.0291233583354044, -0.01198639082531873)

Statistical significance ✓ Practical significance ✓

Net conversion

Pooled probability p\_pool = 0.115127485312

Standard error SE = 0.00343413351293

Margin of error ME = 0.00673090168535

Difference d = -0.00487372267454

Confidential interval CI = (-0.011604624359891718, 0.001857179010803383)

0 ∈ (-0.011604624359891718, 0.001857179010803383) ; dmin = -0.0075 ∈ (-0.011604624359891718, 0.001857179010803383)

Statistical significance ✘ Practical significance ✘

❓ If you have chosen multiple evaluation metrics, you will need to decide whether to use the Bonferroni correction. When deciding, keep in mind the results you are looking for in order to launch the experiment. Will the fact that you have multiple metrics make those results more likely to occur by chance than the alpha level of 0.05?

❗ We have measured two metrics in one experiment. Applying the Bonferroni correction means that the a-level for each hypothesis will be 2.5 % instead of 5% and confidential intervals will be significantly wider. It is too conservative for some reasons.

The use of the Bonferroni correction would really be needed if we test several metrics in one experiment and expect that at least one metrics will demonstrate the statistically significant change. In the set of metrics, this matching only for one indicator can be an absolutely random event, therefore the experiment will have a false positive result. It means we should increase the confidential intervals to avoid this situation and apply the Bonferroni correction.

But in the case of our experiment, we expect two metrics will have matched our criteria at the same time to proceed with the launch. It's a very strong condition without any correction. The positive results will be more likely to occur not by chance. Therefore, the Bonferroni correction could be the cause to approve the wrong null hypothesis and we should not use it this time.

Also, our metrics have a strong relationship between each other. If we know the outcome of one test of a difference between the control and experimental groups on one metrics, it would be easy to predict and to find the outcome of the other tests on the other metrics. It's absolutely natural to expect their behavior will be similar simultaneously.

##### *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.

(These should be the answers from the "Effect Size Tests" quiz.)

#### 🔴   I did not use the Bonferroni correction.

#### 🔵   Gross conversion (Difference between the control and experimental groups)

Lower bound = -0.0291; Upper bound = -0.0120; Statistical significance = ✅; Practical significance = ✅

#### 🔵   Net conversion (Difference between the control and experimental groups)

Lower bound = -0.0116; Upper bound = 0.0019; Statistical significance = ❎; Practical significance = ❎

**Sign Tests**

experiment\_net\_c <- experiment$Payments/experiment$Clicks

experiment\_net\_c <- experiment\_net\_c[!is.na(experiment\_net\_c)]

control\_net\_c <- control$Payments/control$Clicks

control\_net\_c <- control\_net\_c[!is.na(control\_net\_c)]

diff\_net\_c <- experiment\_net\_c - control\_net\_c

avg\_net\_c <- mean(diff\_net\_c)

avg\_net\_c

experiment\_gross\_c <- experiment$Enrollments/experiment$Clicks

experiment\_gross\_c <- experiment\_gross\_c[!is.na(experiment\_gross\_c)]

control\_gross\_c <- control$Enrollments/control$Clicks

control\_gross\_c <- control\_gross\_c[!is.na(control\_gross\_c)]

diff\_gross\_c <- experiment\_gross\_c - control\_gross\_c

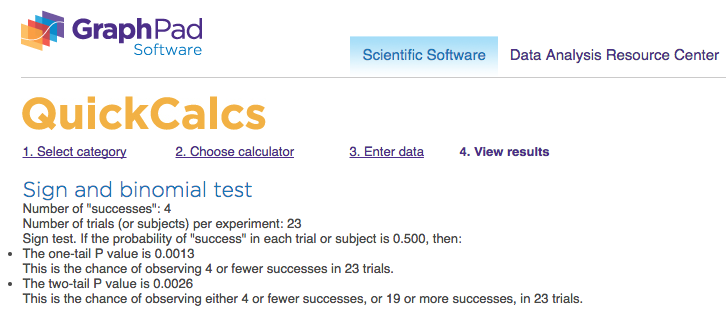
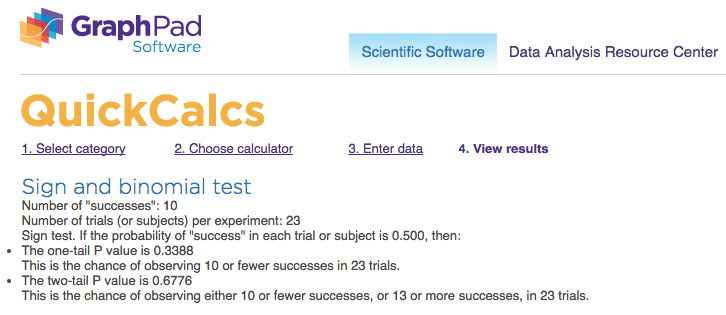
avg\_gross\_c <- mean(diff\_gross\_c)

avg\_gross\_c

Gross conversion: success = 4 total = 23

Net conversion: success = 10 total = 23

 I have used the online calculator (References, N6) for the sign tests. <https://www.graphpad.com/quickcalcs/binomial1.cfm>



For each of your evaluation metrics report the p-value of the sign test and whether the result is statistically significant.

(These should be the answers from the "Sign Tests" quiz.)

#### 🔴   I did not use the Bonferroni correction.

#### 🔵   Gross conversion: p-value = 0.0026; statistical significance = ✅

#### 🔵   Net conversion: p-value = 0.6776; statistical significance = ❎

**Summary**

❓ State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

❗ Eventually, the effective size and sign tests show that the site change would statistically significantly reduce the gross conversion, but would not affect the net conversion in a statistically significant way. The effect size test states this in the practical meaning also.

I did not use the Bonferroni correction for any calculations because it will be too conservative in the case of highly related metrics which should give the expected result simultaneously.

.

**Recommendation**

This experiment was designed to determine whether filtering students as a function of study time commitment would improve the overall student experience and the coaches' capacity to support students who are likely to complete the course, without significantly reducing the number of students who continue past the free trial. A statistically and practically signficant decrease in Gross Conversion was observed but with no significant differences in Net Conversion. This translates to a decrease in enrollment not coupled to an increase in students staying for the requisite 14 days to trigger payment. Considering this, my recomendation is not to launch, but rather to pursue other experiments.

Follow-Up Experiment