index

January 21, 2022

1 A/B Testing - Lab

1.1 Introduction

In this lab, you'll go through the process of designing an experiment.

1.2 Objectives

You will be able to:

- Design, structure, and run an A/B test

1.3 The Scenario

You've been tasked with designing an experiment to test whether a new email template will be more effective for your company's marketing team. The current template has a 5% response rate (with standard deviation .0475), which has outperformed numerous other templates in the past. The company is excited to test the new design that was developed internally but nervous about losing sales if it is not to work out. As a result, they are looking to determine how many individuals they will need to serve the new email template in order to detect a 1% performance increase.

1.4 Step 1: State the Null Hypothesis, H_0

State your null hypothesis here (be sure to make it quantitative as before)

[ ]:
```
# H_0 = Your null hypothesis
# H_0: the probability of success is less than 6 percent, i.e., <0.06
```

1.5 Step 2: State the Alternative Hypothesis, H_1

State your alternative hypothesis here (be sure to make it quantitative as before)

[ ]:
```
# H_1 = Your alternative hypothesis
# H_1: The probability of success is more than 6 percent, i.e., >=0.06
```

1.6 Step 3: Calculate n for standard alpha and power thresholds

Now define what α and β you believe might be appropriate for this scenario. To start, arbitrarily set α to 0.05. From this, calculate the required sample size to detect a .01 response rate difference at a power of .8.

Note: Be sure to calculate a normalized effect size using Cohen's d from the raw response rate difference.

[1]:
```
# Calculate the required sample size
alpha = 0.05
power = 0.8
mean_diff = 0.01
sd = 0.0475

effect_size = mean_diff / sd
```

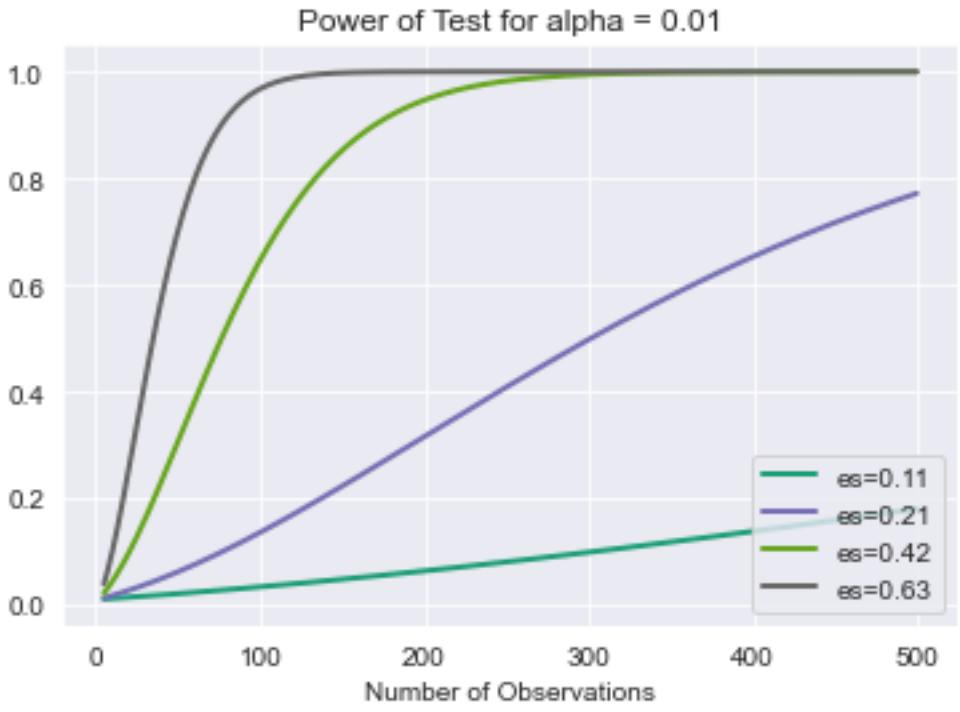
1.7 Step 4: Plot Power Curves for Alternative Experiment Formulations

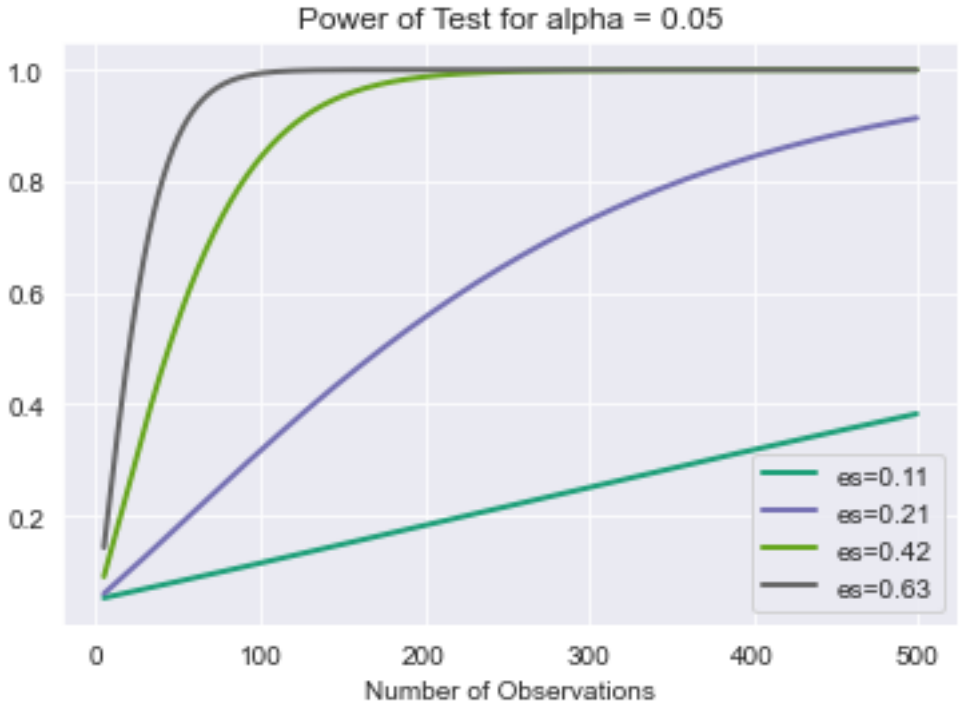
While you now know how many observations you need in order to run a t-test for the given formulation above, it is worth exploring what sample sizes would be required for alternative test formulations. For example, how much does the required sample size increase if you put the more stringent criteria of $\alpha = .01$? Or what is the sample size required to detect a .03 response rate difference at the same α and power thresholds? To investigate this, plot power vs sample size curves for alpha values of .01, .05 and .1 along with varying response rate differences of .005, .01, .02 and .03.

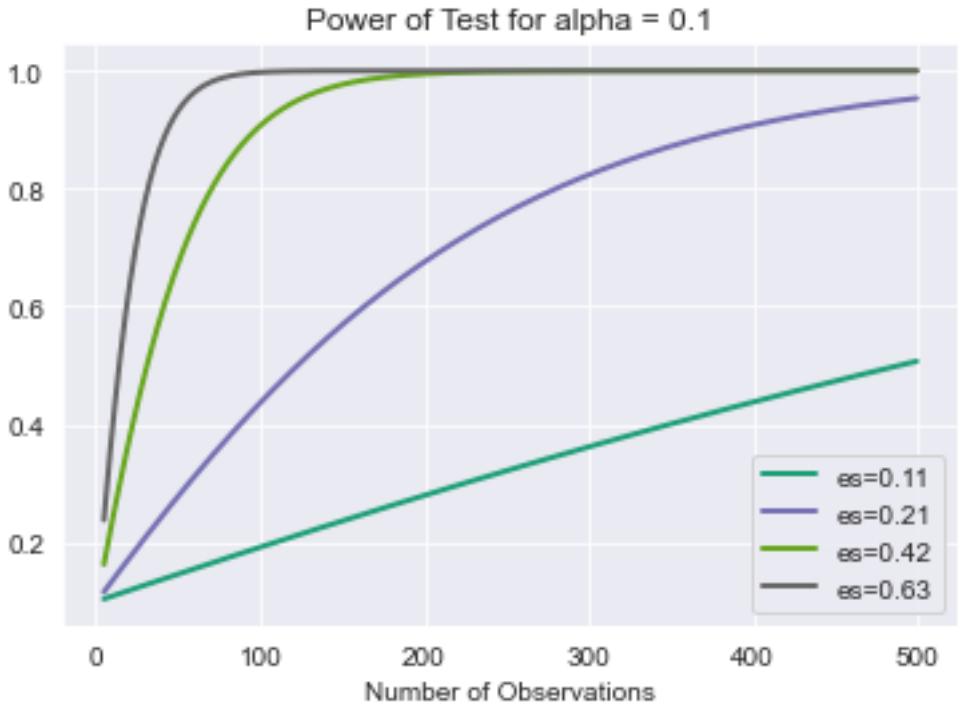
[13]:
```
#Your code; plot power curves for the various alpha and effect size combinations
from statsmodels.stats.power import TTestIndPower

import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
%matplotlib inline
sns.set_style("darkgrid")


power_analysis = TTestIndPower()
```

[14]:
```
alpha = [0.01, 0.05, 0.1]


mean_diff = [0.005, 0.01, 0.02, 0.03]
sd = 0.0475

effect_size = [mean / sd for mean in mean_diff]

for n, item in enumerate(alpha):
    power_analysis.plot_power(dep_var='nobs',
                              nobs = np.array(range(5,500)),
                              effect_size = effect_size,
                             alpha = item)
    plt.title(f"Power of Test for alpha = {item}")
```

```
    plt.show()
```


Power of Test for alpha = 0.01
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Number of Observations
es=0.11
es=0.21
es=0.42
es=0.63


Power of Test for alpha = 0.05
1.0
0.8
0.6
0.4
0.2
0
100
200
300
400
500
Number of Observations
es=0.11
es=0.21
es=0.42
es=0.63



[15]:
```
# ### From GitHub

# #Your code; plot power curves for the various alpha and effect size␣
 ↪combinations
# import numpy as np
# import matplotlib.pyplot as plt
# import seaborn as sns
# sns.set_style('darkgrid')
# %matplotlib inline


# sd = 0.0475
# e_sizes = [mu_delta/sd for mu_delta in [.005,.01,.02,.03]]
# fig, axes = plt.subplots(ncols=1, nrows=3, figsize=(8,15))
# for n, alpha in enumerate([.01, .05, .1]):
#     print(type(n), alpha)
#     ax = axes[n]
#     power_analysis.plot_power(dep_var="nobs",
#                               nobs = np.array(range(5,500)),
#                               effect_size=e_sizes,
#                               alpha=alpha,
#                               ax=ax,
```

```
#                                 alternative='larger')
#     ax.set_title('Power of Test for alpha = {}'.format(alpha))
#     ax.set_xticks(list(range(0,500,25)))
#     ax.set_yticks(np.linspace(0,1,11))
```

1.8 Step 5: Propose a Final Experimental Design

Finally, now that you've explored some of the various sample sizes required for statistical tests of varying power, effect size and type I errors, propose an experimental design to pitch to your boss and some of the accompanying advantages or disadvantages with it.

1.8.1 Your answer here

[23]:
```
alpha = [0.01, 0.05, 0.1]
mean_d = [0.005, 0.01, 0.02, 0.03]
sd = 0.0475

effect_s = [np.round(mean / sd, 2) for mean in mean_d]
sample_size ={}
sample_size_mean_diff ={}
for item in alpha:
    effect_size = {}
    mean_diff = {}
    for n,i in enumerate(effect_s):
        effect_size[i] = power_analysis.solve_power(effect_size=i,
                                                    alpha=item,
                                                    power=.8)
        mean_diff[mean_d[n]] = power_analysis.solve_power(effect_size=i,
                                                    alpha=item,
                                                    power=.8)
    sample_size[item] = effect_size
    sample_size_mean_diff[item] = mean_diff

# sample_size
```

[24]:
```
df_size_effect= pd.DataFrame.from_dict(sample_size)
df_mean_diff= pd.DataFrame.from_dict(sample_size_mean_diff)
```

[25]:
```
df_size_effect.head()
```

[25]:

	0.01	0.05	0.10
0.11	1932.067378	1298.293398	1022.534397
0.21	531.319509	356.920304	281.052392
0.42	134.082656	89.959860	70.779600
0.63	60.530340	40.533933	31.851191

[26]:
```
df_mean_diff.head()
```

[26]:

	0.01	0.05	0.10
0.005	1932.067378	1298.293398	1022.534397
0.010	531.319509	356.920304	281.052392
0.020	134.082656	89.959860	70.779600
0.030	60.530340	40.533933	31.851191

From GitHub Solution

```
Answers will vary. It seems that a minimum sample size 100,
to detect all but the largest effect sizes with a reasonable balance of alpha and power.
After the initial roll-out, there should be sufficient evidence to determine whether
further investigation is warranted.
```

1.9 Summary

In this lab, you practiced designing an initial experiment and then refined the parameters of the experiment based on an initial sample to determine feasibility.