Lecture 5: Randomization and design: The test of significance

In the previous few lectures, we spent time getting a view of experiments as samples from a distribution that has a population or an analytical expression parameterized by the measures. We also learned that randomizing different design variables is a key aspect of accounting for confounding and decreasing the ambiguity in our inferences. In this lecture, we will build on those concepts and try to analyze computationally, the results of an experiment to understand the importance of design in making a particular inference. Because this is a course on the design of experiments, not statistics, our focus is mainly on understanding how a certain design or viewpoint of the data obtained helped us in the final inference we were able to draw. In any experiment, one of the primary goals is to infer whether the changes we experimented with resulted in any significant change to the outcome. Suppose we are studying the effect of a new synthesis recipe of making silica nanoparticles in increasing the yield of a fruit plant as an alternative to the existing fertilizer methods. The silica nanoparticles are coated in a trace amount to some of the seeds selected via a coin toss and shipped to the farmers who followed the same procedure they would otherwise. For simplicity, assume that the farmer obtained a total of 20 seeds and planted them in four rows each containing five seeds. After the seeds were planted, we went to the field and took a complicated image using some advanced technology that does not exist currently to get a precise binary estimate of whether a seed has a coating of nanoparticles or not. From the description, we understand that we randomized the design variable assignment which means we have likely satisfied the i.i.d assumption. After some time, we measure the percentage of fruit that was edible from each one of the seeds and recorded them as follows:

S.No	% edible	label
1	89.7	1
2	81.4	0
3	84.5	1
4	84.8	0
5	87.3	0
6	79.7	1
7	85.1	0
8	91.7	1
9	83.7	1
10	84.5	0
11	84.7	1
12	86.1	1
13	83.2	1
14	91.9	1
15	86.3	1
16	79.3	1
17	82.6	0
18	89.1	1
19	83.7	0
20	88.5	0

Table 1: Example data from the edible fruit example above

where 1 is for seeds with coating and 0 otherwise. We can use our estimates from the sample means to obtain $\bar{y}_1 = 85.93, \bar{y}_0 = 84.73$ thus a gain of 1.19 units for using silica nanoparticles. Now the

question is how do we say whether this method has resulted in any significant increase in the total output of edible fruit? Typically there are three approaches: 1) Use a known reference distribution; 2) Random sample model; 3) Randomized design method.

Known reference distribution

Suppose you have access to data from the previous set of batches *without modification* to the seeds. We can look at the distribution of variation in the percentage of edible fruit between batches of 10 (since we sampled 20 seeds with roughly 10 each with modified vs modified.). If the difference we observed 1.19 is *not* frequent in our computed reference distribution, then we can discredit our **null hypothesis** that the apparent gain in percentage yield is *not statistically significant*. The gain is indeed significant because we observed such change only with a tiny fraction of probability previously. Choosing the reference set to compute the distribution should be done carefully to make sure it clearly represents the without-modification batch used as the control.

Random sample model

In this model, we think of each batch of samples to be a random draw from an underlying distribution. Our null hypothesis, in this case, would be that if the gain is *not* significant and the two batches are indeed from a single distribution. On the basis of the random sampling assumption, we can describe the observations of the percentage yields to be independently drawn. Further, suppose that our hypothesis is that two batch samples are from random populations of the same variance but possibly of different means μ_0, μ_1. We can compute the standard error of the expected gain as follows:

$$
\begin{aligned}
V(\bar{y}_0) = \frac{\sigma^2}{n_0}, \quad V(\bar{y}_1) &= \frac{\sigma^2}{n_1} \\
V(\bar{y}_0 - \bar{y}_1) &= \frac{\sigma^2}{n_0} + \frac{\sigma^2}{n_1} && \text{independently distributed} \\
&= \sigma^2 \left(\frac{1}{n_0} + \frac{1}{n_1} \right)
\end{aligned}
$$

Because the errors are independently distributed, we would expect it to be normal because of the central limit effect with the normal deviate given as:

$$
z = \frac{(\bar{y}_0 - \bar{y}_1) - \delta}{\sigma \sqrt{\left(\frac{1}{n_0} + \frac{1}{n_1} \right)}}
$$

As discussed before, we only have access to a sample from each population we are considering so we estimate them using sample statistics:

$$
t = \frac{(\bar{y}_0 - \bar{y}_1) - \delta}{s \sqrt{\left(\frac{1}{n_0} + \frac{1}{n_1} \right)}}
$$

where the sample variance is estimated using a similar assumption of normality before:

$$
s = \frac{\sum (y_0 - \bar{y}_0)^2 + \sum (y_1 - \bar{y}_1)^2}{n_0 + n_1 - 2}
$$

The value t follows a distribution called the student's t-distribution. The probability of any given t values and the number of degrees of freedom can be found using the following python code:

```
from scipy.stats import t as tdist
print(tdist.sf(t, dof))
```

If the probability from the above calculation is below a certain threshold 5%, we would argue that the modification has some effect as evidenced by the gain in the percentage yield.

Randomized design method

One of the assumptions we had in the previous model about the independence of errors may not hold if there is a small correlation between consecutive readings (also called autocorrelation). This can simply be because of the way measurement is performed. In such cases, we consider that the random assignment we used is just one possibility of multiple such assignments. For example, to assign labels $0, 1$ for twenty samples can be performed in $\frac{20!}{10!10!} = 184756$ ways. Thus if the actual modification does not have a significant effect on the outcome, the observed change would be equally likely for different assignments of the labels while keeping the yields the same. To keep the computational demands minimal, we illustrate this in a different example in the following code:

```
from math import comb

pounds = np.array([29.2, 11.4, 26.6, 23.7, 25.3, 28.5, 14.2, 17.9, 16.5, 21.1, 24.3])
labels = np.array([1,1,0,0,1,0,0,0,1,1,0])
mean_A = np.mean(pounds[labels==1])
mean_B = np.mean(pounds[labels==0])
print('Observed difference : %.2f'%(mean_A-mean_B))
num_possible = comb(len(pounds), 5)
print('Total assignments possible: ', num_possible)

>>> Observed difference : -1.83
>>> Total assignments possible:  462
```

where we have created a data set called pounds with a total of 11 samples and assigned 5 of them the label A and B to the rest. In this case, the total number of possible assignments into groups A and B are given by $\binom{11}{5} = 462$. The following code iterates over all the possible assignments, computes the difference in averages, and plots a histogram shown in Figure 1.

```
import itertools
import seaborn as sns
import matplotlib.pyplot as plt
all_asgs = list(itertools.combinations(np.arange(len(pounds)), 5))
all_diff = []
for asg in all_asgs:
    A = np.asarray(asg)
    B = np.setdiff1d(np.arange(len(pounds)), A)
    yA = pounds[A]
    yB = pounds[B]
    diff = np.mean(yA)-np.mean(yB)
    all_diff.append(diff)
fig, ax = plt.subplots()
sns.histplot(x=all_diff, kde=True)
ax.axvline(mean_A-mean_B, color='tab:red', ls='--')
plt.show()
```

The red dotted vertical lines correspond to our observed value of -1.83 which has a rough probability of $55/462 \approx 12\%$ probability. Therefore we conclude that the null hypothesis that says method A

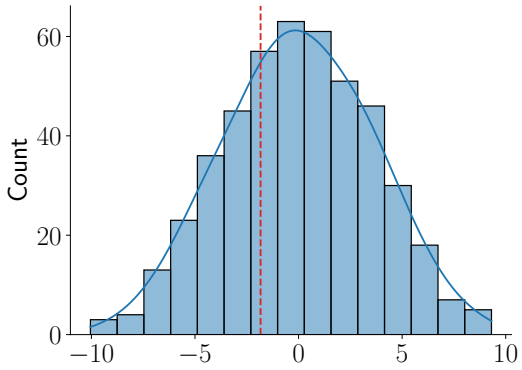

Figure 1: Randomization distribution using the Random design method

is as good as method B cannot be discredited.