P-Hacking

It turns out that science has a bug! If you test many hypotheses but only report the one with the lowest p-value you are more likely to get a spurious result (one resulting from chance, not a real pattern).

Recall p-values: A p-value was meant to represent the probability of a spurious result. It is the chance of seeing a difference in means (or in whichever statistic you are measuring) at least as large as the one observed in the dataset if the two populations were actually identical. A p-value < 0.05 is considered "statistically significant". In class we compared sample means of two populations and calculated p-values. What if we had 5 populations and searched for pairs with a significant p-value? This is called p-hacking!

To explore this idea, we are going to look for patterns in a dataset which is totally random – every value is Uniform(0,1) and independent of every other value. There is clearly no significance in any difference in means in this toy dataset. However, we might find a result which looks statistically significant just by chance. Here is an example of a simulated dataset with 5 random populations, each of which has 20 samples:

	Pop 1	**Pop 2**	**Pop 3**	**Pop 4**	**Pop 5**
1	0.330	0.272	0.959	0.985	0.175
2	0.386	0.353	0.929	0.575	0.386
3	0.232	0.839	0.009	0.229	0.899
	0.836	0.003			
	...				
		0.833	0.333	0.133	
	0.649	0.723	0.565	0.061	0.479
20	0.726	0.158	0.678	0.498	0.645
Sample mean	0.534	0.579	0.474	0.437	0.545

The numbers in the table above are just for demonstration purposes. You should not base your answer off of them. We call each population a random population to emphasize that there is no pattern.

There are Many comparisons

How many ways can you choose a pair of two populations from a set of five to compare? The values of elements within the population do not matter nor does the order of the pair.

$$\binom{5}{2}$$

Understanding the mean of IID Uniforms

What is the variance of a Uniform(0, 1)?

Let $Z \sim \text{Uni}(0,1)$

$$\begin{aligned}\text{Var}(Z) &= \frac{1}{12}(\beta - \alpha) \\ &= \frac{1}{12}(1-0) \\ &= \frac{1}{12}\end{aligned}$$

What is an approximation for the distribution of the mean of 20 samples from Uniform(0,1)?

Let $Z_1 \dots Z_n$ be i.i.d. $\text{Uni}(0,1)$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

$$\text{E}[X] = \frac{1}{n}\sum_{i=1}^{n} E[Z_i] = \frac{1}{n}\sum_{i=1}^{n} 0.5 = \frac{n}{n}0.5 = 0.5$$

$$\begin{aligned}\text{Var}(X) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) \\ &= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} Z_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left(Z_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} v \\ &= \frac{n}{n^2}v = \frac{v}{n} = \frac{v}{20} = \frac{1}{240}\end{aligned}$$

Using CLT, $\bar{X} \sim N\left(\mu = 0.5, \sigma^2 = \frac{1}{240}\right)$

What is an approximation for the distribution of the mean from one population minus the mean from another population? Note: this value may be negative if the first population has a smaller mean than the second.

Let X_1 and X_2 be the means of the populations.
$X_1 \sim N(\mu = 0.5, \sigma^2 = \frac{1}{240})$
$X_2 \sim N(\mu = 0.5, \sigma^2 = \frac{1}{240})$ The expectation is simple to calculate because

$$E[X_1 - X_2] = E[X_1] - E[X_2] = 0$$

$$\begin{aligned}\text{Var}(X_1 - X_2) &= \text{Var}(X_1) + \text{Var}(X_2) \\ &= \frac{1}{120}\end{aligned}$$

The sum (or difference) of independent normals is still normal: \fbox{$Y \sim N(\mu = 0, \sigma^2 = \frac{v}{10})$}

(8 points) What is the smallest difference in means, k, that would look statistically significant if there were only two populations? In other words, the probability of seeing a difference in means of k or greater is < 0.05.

One tricky part of this problem is to recognize the double sidedness to distance. We would consider it a significant distance if $P(Y < -k)$ or $P(Y > k)$.

$$\begin{aligned}P(Y < -k) + P(Y > k) &= 0.05 \\ F_Y(-k) + (1 - F_Y(k)) &= 0.05 \\ (1 - F_Y(k)) + (1 - F_Y(k)) &= 0.05 \\ 2 - 2F_Y(k) &= 0.05 \\ F_Y(k) &= 0.975\end{aligned}$$

Now we need the inverse Φ to get the value of k out.

$$0.975 = \Phi\Big(\frac{k-0}{\sqrt{v/10}}\Big)$$
$$\Phi^{-1}(0.975) = \frac{k}{\sqrt{v/10}}$$
$$k = \Phi^{-1}(0.975)\sqrt{v/10}$$

(5 points) Give an expression for the probability that the smallest sample mean among 5 random populations is less than 0.2.

Let X_i be the sample mean of population i.

$$\begin{aligned}
P(min\{X_1 \dots X_n\} < 0.2) &= P\left(\bigcup_{i=1}^{5} X_i < 0.2\right) \\
&= 1 - P\left(\left(\bigcup_{i=1}^{5} X_i < 0.2\right)^{\complement}\right) \\
&= 1 - P\left(\bigcap_{i=1}^{5} X_i \geq 0.2\right) \\
&= 1 - \prod_{i=1}^{5} P(X_i \geq 0.2) \\
&= 1 - \prod_{i=1}^{5} 1 - \Phi\Big(\frac{0.2-0.5}{\sqrt{v/20}}\Big)
\end{aligned}$$

(7 points) Use the following functions to write code that estimates the probability that among 5 populations you find a difference of means which would be considered significant (using the bootstrapping method designed to compare 2 populations). Run at least 10,000 simulations to estimate your answer. You may use the following helper functions.

```
# the smallest difference in means that would look statistically significant
k = calculate_k()

# create a matrix with n_rows by n_cols elements, each of which is Uni(0, 1)
matrix = random_matrix(n_rows, n_cols)

# from the matrix, return the column (as a list) which has the smallest mean
min_mean_col = get_min_mean_col(matrix)

# from the matrix, return the row (as a list) which has the largest mean
max_mean_col = get_max_mean_col(matrix)

# calculate the p-value between two lists using bootstrapping (like in pset5)
p_value = bootstrap(list1, list2)
```

Write pseudocode:

```
n_significant = 0
k = calculate_k()
for i in range(N_TRIALS):
    dataset = random_matrix(20, 5)
    col_max = get_max_mean_col(dataset)
    col_min = get_min_mean_col(dataset)}
    diff = np.mean(col_max) - np.mean(col_min)}
    if diff >= k:
        n_significant += 1}

print(n_significant / N_TRIALS)
```