One-way Analysis of Variance (ANOVA)

The previous scenarios have been concerned with distinguishing between a sample and a baseline, and between two samples. Suppose you want to distinguish between three or more samples, that is your data fall into three-plus categories and you want to establish whether there is a difference in outcomes based on those categories.

If AAVAIL wanted to run the performance tests from the previous example comparing several cloud providers and several architectures then an ANOVA would be more appropriate here.

```
import numpy as np
from scipy import stats

local_arrivals = np.array([3.99, 4.15, 7.88, 4.53, 5.65, 6.75, 7.13, 3.79, 6.20,
                           3.72, 7.28, 5.23, 4.72, 2.04, 4.25, 4.71, 3.16, 3.46,
                           3.41, 7.98, 0.75, 3.64, 6.25, 6.86, 4.71])
cloud1_arrivals = np.array([5.82, 4.83, 7.19, 6.98, 5.82, 5.25, 5.71, 5.59, 6.93
  ,
                            7.09, 6.37, 6.31, 6.28, 3.12, 6.02, 4.84, 4.16, 6.72
                              ,
                            7.44, 6.28, 6.37, 4.27, 6.15, 4.88, 6.78])
cloud2_arrivals = np.array([5.73, 4.95, 6.96, 6.12, 5.85, 6.74, 5.19, 7.24,
                            6.08, 6.11, 6.11, 7.68, 4.66, 6.12, 5.04, 4.19, 6.46
                              ,
                            7.02, 7.28, 6.19, 4.67, 7.15, 4.58, 6.01])

```

```
all_arrivals = [local_arrivals, cloud1_arrivals, cloud2_arrivals]
global_mean = np.hstack(all_arrivals).mean()

print("The global mean arrival time is: %s"%np.round(global_mean, decimals=2))

for name, arrivals in zip(['local', 'cloud1', 'cloud2'], all_arrivals):
    print("Mean arrival time for {} is {}".format(name, np.round(arrivals.mean
      (), decimals=2)))

```

```
The global mean arrival time is: 5.59
Mean arrival time for local is 4.89
Mean arrival time for cloud1 is 5.89
Mean arrival time for cloud2 is 6.01

```

When comparing across three or more groups (in this case types of promotions) an appropriate test is a one-way ANOVA, which compares between group variation and within group variation. The relevant probability distribution if the F distribution, and that is the name used in the relevant method in Scipy:

```
test_statistic, pvalue = stats.f_oneway(*all_arrivals)
print(np.round(pvalue,decimals=4))

```

```
0.0082
```

In this example, it seems likely that there is at least one difference between the groups. When digging deeper to determine which type of compute environment is best, one needs to be mindful of the multiple comparison problem.