Advanced Practical 2022/2023
Operations Research Case

Lecture: Design of Experiments

Guanlian Xiao

Department Operations Analytics
Vrije Universiteit Amsterdam

June 2023

Purpose

- Consider a stochastic system that is modeled and analysed by discrete event simulation.
- The system is controllable by one or more parameters (or factors).
- Denote by θ the controllable parameter(s).
- Here, we assume that θ takes on a finite number of values, say $\theta \in \Theta = \{\theta_1, \ldots, \theta_d\}$.
- Equivalently, these are alternative system configurations (scenarios) of interest.
- The response or output of the simulation is a performance measure $J(\theta)$ when parameter value or scenario θ is applied.
- We wish to compare (and optimize) the alternative system responses $J(\theta)$ for $\theta \in \Theta$.
- Optimization with respect to continue-valued parameter(s) is the topic of other lecture(s).

Topics

- Comparing two alternatives
  - –> Paired-t confidence interval
  - –> Two-sample t test
  - –> Two-sample Welch's t test
- Comparing more than two systems
  - –> ANOVA
  - –> Pairwise comparison
  - –> Tukey test

Illustration

- A queueing system.
- Poisson arrivals with rate λ p.u.t.
- Two exponential servers at rate μ units p.u.t .
- Service requirement of a customer is either 1 unit with probability p, or 2 units with probability $1 - p$.
- Called type 1 and type 2 customers, respectively.
- These parameters are such that

$$p\lambda < \mu \text{ and } 2(1-p)\lambda < \mu.$$

- Two alternative service policies.

Two Alternative Configurations

A. FCFS with single line queueing. Both servers handle all customers.

This is a $M/H_2/2$ queueing system with load

$$\frac{p\lambda + 2(1-p)\lambda}{2\mu} < 1,$$

thus the queueing system is stable.

B. Server 1 handles only type 1 customers, server 2 the type 2 customers. Upon arrival a customer declares to require 1 or 2 units, and joins the designated queue. This is a system of two (independent) $M/M/1$ queues, both stable.

The performance measure of interest $J(\theta)$ is the expected average waiting time of the type 1 customers (system starts empty) among the first $N = 100$ arrivals, where $\theta \in \{A, B\}$ represent the two scenarios.

The Stochastic Models

- Let X be the random input of the $M/H_2/2$ system.
- It consists of
  - –> the N interarrival times (from exponential distribution with rate λ);
  - –> the N choices of customer type (from Bernoulli distribution with success probability p);
  - –> and the N service durations (from exponential distribution with rate μ or rate $\mu/2$ dependent on the Bernoulli choices).
- The response in system A is the output $W_A = h_A(X) =$ the average waiting of type 1 customers, with expected value $J_A = \mathbb{E}[h_A(X)]$.
- Similarly, the response in system B is the output $W_B = h_B(X) =$ the average waiting of type 1 customers, with expected value $J_B = \mathbb{E}[h_B(X)]$.
- Question is how to evaluate the difference $J_A - J_B$.

DES of the Two System Configurations

- Rather straightforward to execute a run of the first N customers by a discrete event simulation.
- Repeat n times.
- Thus, we generate $X_1, \ldots, X_n$ i.i.d. inputs, and compute the associated average waiting times $W_{A1}, \ldots, W_{An}$ of type 1 customers in system A.
- Independently, generate an additional n i.i.d. inputs $X_{n+1}, \ldots, X_{2n}$, and compute the associated average waiting times $W_{B1}, \ldots, W_{Bn}$ of type 1 customers in system B.

Comparing the Measurements

- Should we compare one-by-one the responses W_{Ai} and W_{Bi}, $i = 1, 2, \ldots$?
- Single observations W_{Ai} are bad estimators of the performance measure J_A, although $\mathbb{E}[W_{Ai}] = J_A$ (unbiased).
- Should we split the observations

$$\underbrace{W_{A1}, \ldots, W_{Am}}_{\text{group } 1}, \underbrace{W_{A,m+1}, \ldots, W_{A,2m}}_{\text{group } 2}, \ldots, \underbrace{W_{A,(k-1)m+1}, \ldots, W_{A,km}}_{\text{group } k},$$

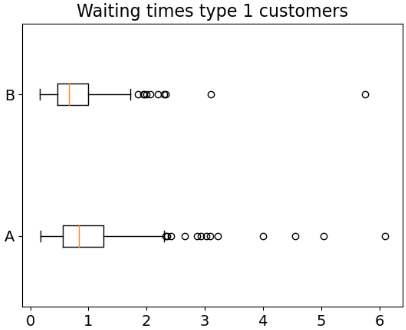
in k groups of size m? Similarly for the W_{Bi} observations. Then compute group averages

$$\overline{W}_{Aj}(m) = \frac{1}{m} \sum_{i=(j-1)m+1}^{jm} W_{Ai}, \quad \overline{W}_{Bj}(m) = \frac{1}{m} \sum_{i=(j-1)m+1}^{jm} W_{Bi},$$

$j = 1, \ldots, k$, and compare those. These group averages are better unbiased estimators of the performnace measure.

Illustration

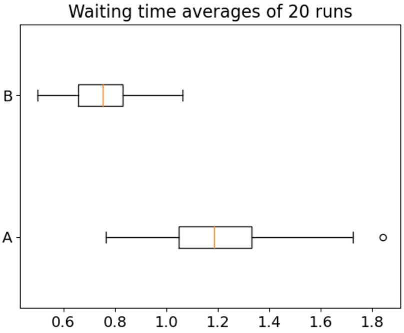
$n = 200$ runs of the two systems. Boxplots of the 200 average waiting times.



It seems that system B has a better performance (smaller waiting time). Difference is small and can be produced by the randomness. Confidence intervals would overlap.

Grouping

$n = 2000$ runs of the two systems. The 2000 observations are split in 100 groups of 20. Boxplots of the 100 group averages of the average waiting times.



Can we now say that system B performs better? We need statistical methods to assess the uncertainty, and to come to decisions.

Paired-t Confidence Interval

- Recall to assess the difference of the expected performances

$$\Delta \doteq J_A - J_B = \mathbb{E}[h_A(X)] - \mathbb{E}[h_B(X)] = \mathbb{E}[h_A(X) - h_B(X)],$$

  where the response function $h_A(X) = W_A$, the average waiting time of type 1 customers in system A.

- The idea is to construct a confidence interval for an estimator of this expectation by running DES to compute the response functions $h_A(X_i)$ and $h_B(X_i)$ for simulated (random) inputs $X_1, \ldots, X_n$.
- Note, that we run the DES for two different systems but with the same input X_i.
- This is called Common Random Numbers (CRN).

CRN

- Using the same random inputs for both systems is called the principle of Common Random Numbers (CRN).
- CRN is an useful technique in comparison and optimization by simulation.
- Using CRN, the measurements (observations, responses) W_{Ai} and W_{Bi} are not independent! This violates the independence assumption that is required in many statistical tests.
- For the paired-t interval method it is crucial.

Constructing the Interval

- Define $Y_i = h_A(X_i) - h_B(X_i)$ ($i = 1, \dots, n$) to be the difference of the computed average waiting times in system A and B when both are fed by the same random input X_i (interarrival times, jobtypes, service requirements).
- The Y_i's are i.i.d., and satisfy $\mathbb{E}[Y_i] = \Delta$.
- Apply the Monte Carlo algorithm: sample average estimator $\overline{Y}(n)$, sample variance $S^2(n)$, estimated standard error $S(n)/\sqrt{n}$ to get a $(1-\alpha)100\%$ confidence interval

$$\overline{Y}(n) \pm t_{n-1,1-\alpha/2}\frac{S(n)}{\sqrt{n}}.$$

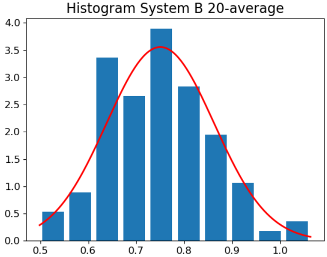
- When the $Y_i's$ are normally distributed, this confidence interval is exact. Otherwise, n should be large.

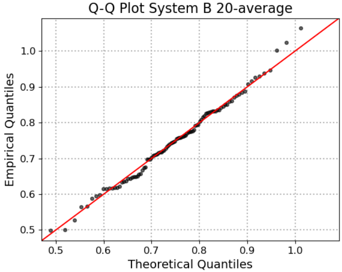
Practice

- Typically in realistic systems, a DES run can take quite some time.
- Thus this limits to execute many runs for obtaining normality.
- However, typically in DES, a response $h(X)$ is the average of of random variables defined over the entire replication, thus might have approximately a normal distribution.
- If not, you might do the grouping idea, meaning that Y_i is itself an avarage of a (small) number of runs to force (approximate) normality.
- For our queueing models, normality tests showed that the average waiting times W_A and W_B are not normal, but the averages $\overline{A}_B(20)$ and $\overline{W}_B(20)$ of 20 runs give sufficient evidence to be approximately normal.

Normality Test

2000 responses W_{Bi} in system B are split in 100 groups of size 20. Thes 20-averages show normality. The p-value of the Shapiro-Wilk test is 79.85%. The histogram and Q-Q plot against fitted normal distribution are plotted below





The Paired-t Interval for Waiting Times

- We simulate 200 runs in both systems using CRN, and split these in 10 groups of size 20.
- Denote $\overline{W}_{Aj}$ and $\overline{W}_{Bj}, j = 1, \ldots, n = 10$ for the average waiting times per group.
- Then, apply the Monte Carlo algorithm for $Y_j = \overline{W}_{Aj} - \overline{W}_{Bj}$.
- We obtain an estimate $\overline{Y}(n) = 0.4095$ with 95% confidence interval $(0.34360.4755)$.
- Conclude that system B performs better for the waiting times of type 1 customers.
- Note, no requirement on the variances of the variables $\overline{W}_A(20)$ and $\overline{W}_B(20)$.

Two-Sample t Test

When it is not possible to run the experiments with CRN, the classic two-sample t test might be applicable. The general theory goes as follows.

- Suppose that there are two normally distributed populations, say N(μ_X, σ_X^2), and N(μ_Y, σ_Y^2).
- The parameters are unknown.
- We wish to test for $\mu_X = \mu_Y$, and $\sigma_X^2 = \sigma_Y^2$.
- Equivalently, we wish to construct confidence intervals for $\mu_X - \mu_Y$, and σ_X^2/σ_Y^2.
- We have i.i.d. samples $X_1, \ldots, X_m$ of X, and $Y_1, \ldots, Y_n$ of Y (allowed is $m \neq n$). And also the two sample sets are independent of each other!

Two-Sample t Test for the Difference

- Assume that the variance $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ of the populations are equal. See next topic (Welch's t test fro dropping this assumption.
- Compute sample mean and sample variance of the samples, $\overline{X}, S_X^2, \overline{Y}, S_Y^2$, as usual.
- Compute the pooled sample variance

  $$S_P^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}.$$

  This is an unbiased estimator of σ^2.
- Then, the test statistic for the hypothesis $H_0 : \mu_X - \mu_Y = \Delta$ (for instance $\Delta = 0$) is

  $$T = \frac{\overline{X} - \overline{Y} - \Delta}{\sqrt{S_P^2/m + S_P^2/n}},$$

  which has a t-distribution with $m+n-2$ degrees of freedom.
- Equivalently, a $(1-\alpha)100\%$ confidence interval for $\mu_X - \mu_Y$ is

  $$\overline{X} - \overline{Y} \pm t_{m+n-2,1-\alpha/2}\, S_P \sqrt{\frac{1}{m} + \frac{1}{n}}.$$

Confidence Interval for the Ratio

- You might need to test for equal variances.
- It holds that the pivot

$$T = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

  has a F distribution with $m-1$ and $n-1$ degrees of freedom.
- Then a $(1-\alpha)100\%$ confidence interval for σ_X^2/σ_Y^2 is

$$\left(\frac{1}{F_{m-1,n-1,1-\alpha/2}}\frac{S_X^2}{S_Y^2}, \frac{1}{F_{m-1,n-1,\alpha/2}}\frac{S_X^2}{S_Y^2}\right).$$

- Alternative: perform the Levene test, or the Bartlett test. These are available in `scipy.stats`.

Application to the Queueing Systems

- We simulate 200 runs in both systems, independently, and split these in $n = 10$ groups of size 20.
- We form again the averages of the groups because we saw that we get samples from normally distributed populations,

$$\overline{W_{A1}}, \ldots, \overline{W_{An}} \overset{\text{iid}}{\sim} \mathsf{N}(J_A, \sigma_A^2).$$

- The computed 95% confidence interval for the ratio σ_A^2/σ_B^2, based on the experiments is $(0.4139, 6.7093)$.
- We see that 1 is an element of the interval, thus we may assume that $\sigma_A^2 \approx \sigma_B^2$.
- Next, compute the pooled sample variance, and the test statistic, we get $T = 4.9814$, p-value $= 9.66\ 10^{-5} \approx 0$, and 95% confidence interval for $J_A - J_B$, $(0.1609, 0.3958)$.
- Again we see that system B performs better.

Two-Sample Welch's t Test

- The same problem as the two-sample t test, without assuming equal variances.
- Refer to the notation of slides 17-18.
- The test statistic for the hypotheis $H_0 : \mu_X = \mu_Y$ is

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2/m + S_Y^2/n}},$$

which has a t-distribution with ν degrees of freedom, which is estimated by

$$\hat{\nu} = \frac{\left(S_X^2/m + S_Y^2/n\right)^2}{\left(S_X^2/m\right)^2/(m-1) + \left(S_Y^2/n\right)^2/(n-1)}.$$

- Equivalently, a $(1-\alpha)100\%$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm t_{\hat{\nu},1-\alpha/2}\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}.$$

Application to the Queueing Systems

- We simulate 200 runs in both systems, independently, and split these in $n = 10$ groups of size 20 for normality.
- Computing the sample variances, and the test statistic, we get $T = 4.9814$, p-value $= 1.15\ 10^{-4} \approx 0$, and 95% confidence interval $(0.1604, 0.3963)$.
- Nearly the same as the standard two-sample t test assuming equal variances.

Python

- The function `ttest_ind` is available in `scipy.stats` and in `statsmodels.stats.weightstats`.
- You can specify whether variances are assumed to be equal or not.
- In the latter case Welch's t test is performed.
- The function returns the test statistic and the p-value for the hypothesis of equal means.
- The `statsmodels` version returns also the degrees of freedom.

Comparing More Than Two Systems

Illustration

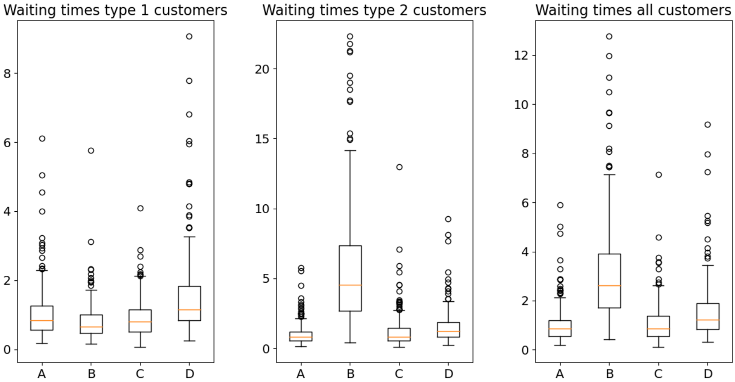
Consider again the queueing system with the configurations A and B. Now add two more systems, receiving the same random input of arrival Poisson process, Bernoulli jobtypes, and service requirements. All systems will have the same work load.

- *C.* Same as B, but now, when server 1 becomes free while no type 1 customers are present but there are type 2 customers waiting, he starts serving a type 2. Similarly concerning server 2.
- *D.* The queue $M/H_2/1$ with one server who works twice as fast as the 2 servers in $M/H_2/2$. Service is FCFS no matter the jobtype.

Four systems to compare. For instance, average waiting time of type 1 cutomers, type 2 customers, or all customers. See next slide.

NB, these are not long-run averages (would require steady-state simulation), but the averages of the first 100 customers. Each system is replicated 200 times independently.

Boxplots


Waiting times type 1 customers
Waiting times type 2 customers
Waiting times all customers
0
2
4
6
8
5
10
15
20
12
A
B
C
D

Confidence Intervals

- We have four scenarios, A, B, C, D.
- Consider the expected average waiting time of type 1 customers as the performance m easure of interest, denote J_A, J_B, J_C, J_D.
- We wish to construct confidence intervals for all 6 difference pairs,

  $$\Delta_1 = J_A - J_B, \Delta_2 = J_A - J_C, \Delta_3 = J_A - J_D, \Delta_4 = J_B - J_C, \Delta_5 = J_B - J_D, \Delta_6 = J_C - J_D.$$

- Let $I_k, k = 1, \dots, 6$ be these intervals. We wish these such that

  $$\mathbb{P}(\Delta_k \in I_k, k = 1, \dots, 6) \geq 1 - \alpha.$$

  Then the intervals I_k should be constructed such that $\mathbb{P}(\Delta_k \in I_k) \geq 1 - \alpha/6$ for all $k = 1, \dots, 6$.
- This is based on a Bonferroni inequality, see next slide.

Bonferroni Inequality

- Recall the Monte Carlo algorithm for estimating a performance measure μ (expected value of the response fiunction).
- An unbiased estimator $\widehat{Y}$ is implemented (e.g. sample average) with standard deviation σ (also known as standard error).
- Standard error is usually estimated as well (via sample variance), say by $\widehat{\sigma}$.
- Then, a random interval I is constructed, called confidence interval, by $I = (\widehat{Y} - t\widehat{\sigma}, \widehat{Y} + t\widehat{\sigma})$ such that $\mathbb{P}(I \ni \mu) \geq 1 - \alpha$.
- The confidence level is typically $\alpha = 0.05$. The critical value t depends on α and on normality properties of the estimator. Typically it is a t-distribution quantile.
- Now, suppose to estimate d performance measures $\mu_1, \ldots, \mu_d$. The k-th is estimated such that an associated confidence intervals I_k is constructed with confidence level α_k.
- Then a Bonferrroni inequality says

$$\mathbb{P}\mathbb{P}(I_k \ni \mu_k, k = 1, \ldots, d) \geq 1 - \sum_{k=1}^{d} \alpha_k.$$

Application to Queueing Configurations

- We apply both the paired t method (using CRN) and the two-sample Welch's method to all 6 differences.
- As before, we run all systems 200 times, and compute the 10 averages of groups of size 20.
- As confidence level (overall) we set $\alpha = 0.1$.
- Each of the 6 differences uses confidence $\alpha/6 = 0.01667$.

The Tables

Paired-t tests

	B	C	D
A	$(0.324, 0.495)$	$(0.130, 0.304)$	$(-0.486, -0.167)$
B		$(-0.268, -0.117)$	$(-0.837, -0.635)$
C			$(-0.646, -0.441)$

Welch's t tests

	B	C	D
A	$(0.130, 0.427)$	$(0.009, 0.300)$	$(-0.841, -0.133)$
B		$(-0.248, 0.000)$	$(-1.115, -0.415)$
C			$(-0.991, -0.291)$

All intervals in both methods are significant!

Python: Tukey's Test

- The function `pairwise_tukeyhsd` in `statsmodels.stats.multicomp` is available.
- You specify the data and the scenario labels (as arrays).
- Applied to our 4 queueing systems the following output is returned.

```
Multiple Comparison of Means - Tukey HSD, FWER=0.10
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
     A      B  -0.2784 0.0293  -0.505 -0.0518   True
     A      C  -0.1545 0.3812 -0.3811  0.0721  False
     A      D   0.4866  0.001    0.26  0.7133   True
     B      C   0.1239 0.5626 -0.1028  0.3505  False
     B      D    0.765  0.001  0.5384  0.9916   True
     C      D   0.6412  0.001  0.4145  0.8678   True
----------------------------------------------------
```

Two comparisons are not significant in this approach.
The confidence intervals are slightly different.
This could be due to the assumption of equal variances (for Tukey test).

ANOVA

- Before you run the multiple comparison tests, it is advised to test whether there will be unequal means.
- The one-way (or one-factor) ANOVA is suitable for this (Analysis of Variance).
- The general theory goes as follows.
- Suppose that there are d normally distributed populations (or groups), say $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2), i = 1, \ldots, d$.
- The parameters are unknown.
- We wish to test for all equal means $\mu_1 = \mu_2 = \cdots = \mu_d$, against that at least one is significantly different from the others.
- The data consists of d independent samples of sizes $n_1, \ldots, n_d$,

$$\underbrace{(Y_{11}, \ldots, Y_{1n_1})}_{\overset{\text{iid}}{\sim} Y_1}, \underbrace{(Y_{21}, \ldots, Y_{2n_2})}_{\overset{\text{iid}}{\sim} Y_2}, \ldots, \underbrace{(Y_{d1}, \ldots, Y_{dn_d})}_{\overset{\text{iid}}{\sim} Y_d}.$$

Typically, one says that Y_{ij} is the j-th response variable of group i.

ANOVA (cont'd)

- The third assumption is that all population variances are the same $\sigma_i^2 \equiv \sigma^2$.
- Thus, we get the ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \dots, d; j = 1, \dots, n_i,$$

  where $\{\epsilon_{ij}\} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$.
- Equivalently, write $\mu_i = \mu + \alpha_i$ with $\sum_{i=1}^{d} \alpha_i = 0$.
- In this way is ANOVA a normal linear model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- See your Statistics textbook.
- The hypothesis is $H_0 : \alpha_1 = \cdots = \alpha_d = 0$.

ANOVA (cont'd)

- The test statistic is

$$T = \frac{MS_B}{MS_W} = \frac{\text{mean square between samples}}{\text{mean square within sample}}$$
$$\frac{\text{sum of squares between samples}/(d-1)}{\text{sum of squares within sample}/(n-d)},$$

  where $n = \sum_{i=1}^{d} n_i$ is the total number of observations.

- Under H_0, T has an F distribution with degrees of freedom $d-1$ and $n-d$.
- When the null hypothesis is true, MS_B and MS_W are both estimates of error variance and would be about the same size. Their ratio, would be close to one. When the null hypothesis is not true then the MS_B will be larger than MS_W and their ratio greater than 1.
- Commonly one calls MS_B the mean square of the factor, and MS_W the mean square of the error. Their ratio is called F ratio.
- In order to tell if the F statistic is statistically significant, you look up the critical value based on the degrees of freedom and the confidence level α. Or you compute the associated p-value.

ANOVA (cont'd)

- The sum of squares between samples is computed by

$$SS_B = \sum_{i=1}^{d} n_i \left(\overline{Y}_i - \overline{Y}\right)^2,$$

where $\overline{Y}_i$ is the sample average of the i-th sample, and $\overline{Y}$ is the overal average,

$$\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}; \quad \overline{Y} = \frac{1}{\sum_{i=1}^{d} n_i} \sum_{i=1}^{d} \sum_{j=1}^{n_i} Y_{ij}.$$

The latter is equal to $\frac{1}{d} \sum_{i=1}^{d} \overline{Y}_i$ if all sample sizes n_i are equal.

- The sum of squares between samples is computed by

$$SS_W = \sum_{i=1}^{d} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_i\right)^2 = \sum_{i=1}^{d} (n_i - 1) S_i^2,$$

where S_i^2 is the sample variance of group i.

ANOVA for the Queueing Configurations

- As before, we run all systems 200 times independently, and compute the 10 averages of groups of size 20.
- The three requirements of ANOVA are (i) independence, (ii) normality, (iii) equal variances.
- (i) is satisfied by the simulation procedure. For now, assume (ii) and (iii) hold.
- We get the following tables of computations.

scenario	sample size	sample average	sample variance
A	10	1.075	0.0195
B	10	0.797	0.0117
C	10	0.921	0.0105
D	10	1.562	0.1402
total	40	1.089	

source	sum sqrs	df	mean sq	F statistic	p-value
between samples	3.375	3	1.124	24.74	$7.25\,10^{-9}$
within samples	1.636	36	0.0454		

Conclude H_0 is rejected.

Python

1. The function `f_oneway` in `scipy.stats`.
   - –> You specify the samples (groups) as 4 arrays.
   - –> It returns the test statistic and p-value.
   - –> Applied to our queueing problem, with as groups the four systems, and in each group 10 data values, it returns statistic 24.74 and p-value $7.25\ 10^{-9}$.

2. The function `anova` in the package `pingouin`. This is an open-source statistical package that is not standard included in Anaconda, but easily installable.
   - –> You specify the data in a `pandas` dataframe.
   - –> You specify what column contains the dependent variable, and what column(s) the data with the between-subject factor(s).
   - –> It returns the complete table of the previous slide!

Checking the Assumptions

- The normality is tested by Shapiro-Wilk or Jarque-Bera.
- Using the 200 data grouped in 10 of size 20, Shapiro-Wilk test gave

scenario	test statistic	p-value
A	0.9061	0.2553
B	0.9499	0.6669
C	0.9266	0.4149
D	0.9391	0.5429

  Conclude that all groups may be considered to be normal distributed.
- The equal variance assumption is checked by the Levene of Bartlett test. Levene gave a p-value of $0.014 < 0.05$, thus the hypothesis of equal variances is rejected.
- Then an alternative is to perform the Welch's ANOVA. This is available in package `pingouin`. It finds F statistic equal to 16.99 with degrees of freedom 3 and 19.28 and p-value $1.2\,10^{-5}$.

ANOVA Post-Hoc Analysis

- Once your ANOVA concluded that the sample means differ, you perform post-hoc analysis to identify which particular differences between pairs of means are significant.
- In fact, we have seen these tests above (pair-wise comparisons, and the Tukey test).