

Харківський національний університет ім. В. Н. Каразіна
Факультет комп'ютерних наук
Кафедра теоретичної та практичної системотехніки

УХВАЛЕНО

Вченою радою факультету
комп'ютерних наук, протокол № _____
від «__» _____ 2020 р.

Голова Вченої ради _____



| | |
|---|---|
| Назва курсу | Аналіз даних |
| Викладач (-і) | доцент Мазорчук М.С. |
| Профайл викладача (-ів) | http://www-csd.univer.kharkov.ua/about-us/sub-faculty/tase-department/personalnij-sklad/ |
| Контактний тел. | |
| E-mail: | mazorchuk.mary@gmail.com |
| Сторінка курсу в системі дистанційного навчання | - |
| Консультації | Очні консультації: 6 годин, кожен парний понеділок з 15.20 до 16.20, ауд. 320 |

1. Коротка анотація до курсу - курс «Аналіз даних» спрямований на засвоєння студентами основ моделювання чисельних і категоріальних даних та методології статистичного аналізу, статистичного оцінювання взаємозв'язку величин, зниження розмірності

даних та ін., вироблення навичок по адаптації стандартних алгоритмів до нових – чисельних рішень складних прикладних задач, а також придбання знань про пакети прикладних програм спеціального призначення і середовищ для обробки даних.

2. Мета та цілі курсу – оволодіння системним підходом до аналізу даних, засвоєння методів формального описання даних різної природи, закріплення знань та розвиток практичних навичок аналізу даних за допомогою комп'ютерних систем.

3. Формат курсу - очний

4. Результати навчання –

Відповідно до вимог освітньо-кваліфікаційного рівня підготовки за результатами вивчення дисципліни студенти повинні –

- знати методи статистичного аналізу даних в обсязі, необхідному для користування математичним апаратом та методами у галузі автоматизації, а саме:
- основні цілі та вихідні передумови застосування статистичних методів при управлінні складними комп'ютеризованими системами;
- методи попередньої обробки даних;
- основні поняття вибіркового методу;
- методи перевірки статистичних гіпотез;
- методи кореляційного та регресійного аналізу статистичних даних;

вміти:

- застосовувати сучасні інформаційні технології та мати навички розробляти алгоритми та комп'ютерні програми з використанням мов високого рівня для аналізу даних, створювати бази даних та використовувати інтернет-ресурси;
- вибирати адекватні методи статистичного аналізу даних у відповідності з метою дослідження та характером статистичних даних;
- знаходити рівняння лінійної та нелінійної регресії;
- перевіряти значущість коефіцієнтів лінійної регресії;
- визначати статистичні оцінки для параметрів лінійної та нелінійної регресії;
- представляти змістовну інтерпретацію результатів статистичного аналізу;
- працювати з сучасними програмними системами і середовищами для статистичного аналізу даних (STATISTICA, SPSS, R, пакети Python);

5. Обсяг курсу

| Вид заняття | Загальна кількість годин |
|-------------------|--------------------------|
| лекції | 24 |
| практичні | 24 |
| самостійна робота | 42 |

6. Ознаки курсу:

| Рік викладання | семестр | спеціальність | Курс (рік навчання) | Нормативний\ вибірковий |
|----------------|---------|---|------------------------|----------------------------|
| 2020 | 7 | Автоматизація та комп'ютерно-інтегровані технології | 4 | вибірковий (В) |

7. Пререквізити – перед вивченням курсу студент повинен знати вищу математику, основи теорії ймовірностей та математичної статистики, основи системотехніки, основи філософії, теорії алгоритмів.

8. Технічне та програмне забезпечення /обладнання - STATISTICA, SPSS, R-Studio, пакети Python.

9. Політики курсу

Академічна доброчесність - це сукупність етичних принципів та визначених законом правил, якими мають керуватися учасники освітнього процесу під час навчання, викладання та провадження наукової (творчої) діяльності з метою забезпечення довіри до результатів навчання та/або наукових (творчих) досягнень. Дотримання академічної доброчесності здобувачами освіти передбачає самостійне виконання навчальних завдань, завдань поточного та підсумкового контролю результатів навчання; посилення на джерела інформації у разі використання ідей, розробок, тверджень, відомостей.

10. Схема курсу

| Тиж. / акад.год. | Тема, план, короткі тези | Форма діяльності (заняття)* *лекція, ПЗ,ЛР, СРС) / Формат** **аудиторна, СРС | Матеріали | Завдання, год |
|---|--|---|--|--|
| Розділ 1. Оцінювання невизначеностей при вимірюванні в статистичному дослідженні | | | | |
| | <p><i>Тема 1.</i> Тема 1. Аналіз даних у системах різної природи на основі моніторингу.</p> <p><i>Лекція 1.</i> Застосування математичних методів у різних системах: опис, пояснення та прогнозування статистичних процесів та явищ. Основні напрямки використання математичних методів при аналізі даних.</p> | <i>Лекція / аудиторна</i> | Презентація лекції | 2 год. |
| Тиж. 1 9 год. | <p>СРС. Інформаційно-аналітичне забезпечення процесів діагностування систем різної природи на основі даних моніторингу.</p> | <i>СРС / СРС</i> | <p>Конспект лекцій, Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. – К.: МОРИОН, 2002. – 640 с., Айвазян С.А., Мхитарян В.С. Прикладная статистика и</p> | <p>опрацювати лекцію, ознайомитись з літературою, 5 год.</p> |

| | | | | |
|------------------|---|-----------------------|--|--|
| | | | основы эконометрии. – М.: ЮНИТИ, 1998. – 1000 с. | |
| | ПЗ 1. Знайомство з середовищем R Studio або SPSS Первинна обробка даних | Практичне заняття | Анализ данных результатов статистических исследований в среде R [Текст] : учеб. пособие / М. С. Мазорчук, В. С. Добряк, Н.С. Бакуменко. – Х.: Нац. аэрокосм. ун-т им. Н.Е. Жуковского «Харьк. авиац. ин- т», 2018. – 36 с., Завдання на ПЗ 1, Field, Andy (2009). Discovering Statistics Using SPSS. – 855 p., Field, Andy (2012). Discovering Statistics Using R. – 995 p. | опрацювати лекцію, ознайомитись з літературою, 2 год |
| Тиж. 2 9 год. | Тема 2. Статистичне оцінювання параметрів розподілів. Лекція 2. Вибірка та популяція. Статистичні оцінки. Міри центральної тенденції. Міри варіації. | Лекція / аудиторна | Презентація лекції | 2 год |

| | | | | |
|--|--|--------------------------------------|---|--|
| | <p>СРС. Форми обліку результатів спостереження. Типи та види оцінювання.</p> | <p><i>СРС</i></p> | <p>Конспект лекцій, Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. – К.: МОРИОН, 2002. – 640 с.,</p> | <p>опрацювати лекцію, ознайомитись з літературою, 5 год</p> |
| | <p>ПЗ 2. Перевірка гіпотези про вид розподілу ознаки у середовищі R Studio.</p> | <p>Практичне заняття.</p> | <p>Конспект лекції, завдання на ПЗ 2</p> | <p>опрацювати лекцію, ознайомитись з літературою, 2 год.</p> |
| <p>Тиж. 3 8 год.</p> | <p><i>Тема 3.</i> Кореляційний аналіз даних та оцінка зв'язку між різними типами даних. <i>Лекція 3.</i> Функціональний та кореляційний зв'язки. Діаграма розсіяння. Лінія регресії. Основні властивості статистичної залежності: тип, форма та тіснота зв'язку.</p> | <p><i>Лекція / аудиторна</i></p> | <p>Презентація лекції</p> | <p>2 год.</p> |
| | <p>СРС. Коефіцієнти рангової кореляції Кендела: правила обчислення, область значень та інтерпретація. Порівняння коефіцієнтів Спірмена та Кендела.</p> | <p><i>СРС</i></p> | <p>Конспект лекції, Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. – К.: МОРИОН, 2002. – 640 с. (С 151-180)</p> | <p>опрацювати лекцію, ознайомитись з літературою, 4 год.</p> |

| | | | | |
|--------------------------------|---|---------------------------|---|---|
| | ПЗ 3. Статистичне оцінювання параметрів кореляції ознак у середовищі R Studio. | Практичне заняття | Конспект лекції, завдання на ПЗ 3 | опрацювати лекцію, ознайомитись з літературою, 2 год. |
| Тиж. 4 8 год. | <i>Тема 4.</i> Аналіз категоріальних даних. <i>Лекція 4.</i> Двовимірні таблиці як інструмент вивчення взаємозв'язку двох ознак: структура та правила читання. | <i>Лекція / аудиторна</i> | Презентація лекції | 2 год. |
| | СРС. Коефіцієнти зв'язку для двох ознак, побудовані на основі χ^2 . | <i>СРС</i> | Лапач С.Н., Чубенко А.В., Бабиць П.Н. Статистика в науці та бізнесі. – К.: МОРИОН, 2002. – 640 с. (с. 243-273) | Опрацювати лекцію, ознайомитись з літературою, 4 год. |
| | ПЗ 4. Аналіз категоріальних даних у середовищі R Studio або SPSS. | Практичне заняття. | Конспект лекції, завдання на ПЗ 4, Статистический анализ данных в социологии [Текст] : учеб. пособие / М. С. Мазорчук, В. С. Добряк. – Х. : Нац. аэрокосм. ун-т им. Н.Е. Жуковского «Харьк. авиац. ин-т», 2012. – 36 с. | опрацювати лекцію, ознайомитись з літературою, 2 год. |

| | | | | |
|-----------------|---|----------------------------------|---|--|
| Тиж. 5 8 год | <p><i>Тема 4. Кореляційний аналіз даних. Лінійний та нелінійний зв'язки. Прямий та зворотній зв'язки. Коефіцієнт детермінації як універсальна характеристика ступеню тісноти статистичного зв'язку.</i></p> | <p><i>Лекція / аудиторна</i></p> | <p>Презентація лекції</p> | <p>2 год.</p> |
| | <p>СРС. Дисперсійний аналіз даних (однофакторний та двофакторний)</p> | <p><i>СРС</i></p> | <p>Конспект лекції</p> | <p>опрацювати лекцію, ознайомитись з літературою, 4 год.</p> |
| | <p>ПЗ 5. Дослідження лінійної та нелінійної залежності за допомогою у середовищі R Studio або SPSS.</p> | <p>Практичне заняття.</p> | <p>Конспект лекції, завдання на ПЗ 5, Аналіз даних результатів статистических исследований в среде R [Текст] : учеб. пособие / М. С. Мазорчук, В. С. Добряк, Н.С. Бакуменко. – Х.: Нац. аэрокосм. ун-т им. Н.Е. Жуковского «Харьк. авиац. ин-т», 2018. – 36 с.,</p> | <p>опрацювати лекцію, ознайомитись з літературою, 2 год.</p> |

| | | | | |
|---|---|---------------------------|--|---|
| Тиж. 6 8 год | Лекція 6. Кореляційне відношення. Множинні та часткові коефіцієнти кореляції. Рангова кореляція | <i>Лекція / аудиторна</i> | Презентація лекції | 2 год. |
| | СРС. Конкордація ознак. | <i>СРС</i> | Конспект лекції | опрацювати лекцію, ознайомитись з літературою, 2 год |
| | ПЗ 6. Кореляційний аналіз даних у середовищі R Studio або SPSS. | Практичне заняття. | Конспект лекції, завдання на ПЗ 6, Аналіз даних результатів статистических досліджень в середі R [Текст] : учеб. пособие / М. С. Мазорчук, В. С. Добряк, Н.С. Бакуменко. – Х.: Нац. аэрокосм. ун-т им. Н.Е. Жуковского «Харьк. авиац. ин-т», 2018. – 36 с. | опрацювати лекцію, ознайомитись з літературою, 2 год. |
| Розділ 2. Формальні математичні моделі систем і процесів | | | | |
| Тиж. 7 8 год | <i>Тема 5.</i> Множинний регресійний аналіз. Лекція 7. Регресійний аналіз. Модель лінійної регресії. Обмеження регресійної моделі - мультиколінеарность, гомоскедастичность. | <i>Лекція / аудиторна</i> | Презентація лекції | 2 год |

| | | | | |
|-----------------|---|--------------------|---|--|
| | СРС. Основні методи побудови регресійних моделей. | СРС | Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрии. – М.: ЮНИТИ, 1998. – 1000 с. (с. 455-461) | опрацювати лекцію, ознайомитись із літературою, 4 год |
| | ПЗ 7 Побудова регресійних моделей у середовищі R Studio або SPSS. | Практичне заняття | Конспект лекції, завдання на ПЗ 7, Field, Andy (2009). Discovering Statistics Using SPSS. – 855 p., Field, Andy (2012). Discovering Statistics Using R. – 995 p. | опрацювати лекцію, ознайомитись з літературою, 2 год. |
| Тиж. 8 8 год | Тема 5. Множинний регресійний аналіз. Лекція 8. Модель логістичної регресії. Інтерпретація коефіцієнтів логістичної регресії. Некоректно поставлені завдання. Алгоритми, що регуляризують (робастні алгоритми): адаптивні, інваріантні. | Лекція / аудиторна | Презентація лекції | 2 год. |
| | СРС. Методи регуляризації в задачах ідентифікації, апроксимації даних та прогнозування часових рядів. | СРС | Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы | опрацювати лекцію, ознайомитись із літературою, 4 год. |

| | | | | |
|-------------------------------|---|---------------------------|---|---|
| | | | економетрії. – М.: ЮНІТИ, 1998. – 1000 с. | |
| | ПЗ 8. Множинний регресійний аналіз (лінійна форма регресійної залежності) у середовищі R Studio або SPSS. | Практичне заняття. | Конспект лекції | опрацювати лекцію, ознайомитись з літературою, 2 год. |
| Тиж. 9 8 год | <i>Тема 6.</i> Методи оцінювання інформативності (значущості) змінних при невизначеності даних. <i>Лекція 9.</i> Методи оцінюванні диференціальної інформативності з врахування точності вимірювання змінних стану и наявність парної кореляції между ними: кореляційного аналізу, дисперсійного аналізу і методи розпізнавання образів. Статистичні оцінки довірчих інтервалів для математичного очікування нелінійних залежностей методом Монте-Карло. | <i>Лекція / аудиторна</i> | Презентація лекції | 2 год |
| | СРС. Ймовірносно-статистичні (методи Байеса, послідовного аналізу і оцінювання на основі теорії інформації) методи оцінювання інформативності. | <i>СРС</i> | Конспект лекції, Черняк О.І., Захарченко П.В Інтелектуальний аналіз даних: К., 2014. — 599 с. | опрацювати лекцію, ознайомитись із літературою, 4 год |
| | ПЗ 9. Множинний регресійний аналіз (поліноміальна форма регресійної залежності) у середовищі R Studio. | Практичне заняття | Конспект лекції, завдання на ПЗ 9, Field, Andy (2009). Discovering Statistics Using SPSS. – 855 p., | опрацювати лекцію, ознайомитись з літературою, 2 год. |

| | | | | |
|--------------------------------|--|---------------------------|--|---|
| | | | Field, Andy (2012). Discovering Statistics Using R. – 995 p. | |
| Тиж. 10 7 год | Лекція 10. Оцінювання інформативності на основі методів структурно-параметричного аналізу і синтезу регресійних моделей. | <i>Лекція / аудиторна</i> | Презентація лекції | опрацювати лекцію, переглянути презентацію, 2 год |
| | СРС. Методи послідовного аналізу варіантів, вагові з адаптацією, локально-стохастичні на основі самоорганізації. | <i>СРС</i> | Конспект лекції, Черняк О.І., Захарченко П.В Інтелектуальний аналіз даних: К., 2014. — 599 с. | опрацювати лекцію, ознайомитись із літературою, 4 год |
| | ПЗ 10. Оцінювання інформативності (значущості) змінних при невизначеності даних. | Практичне заняття | Конспект лекції, завдання на ПЗ 10 | опрацювати лекцію, ознайомитись з літературою, 2 год. |
| Тиж. 11 8 год | <i>Тема 7.</i> Факторний аналіз. <i>Лекція 11.</i> Головна ідея методів факторного аналізу даних. Задачі, що розв'язуються із застосуванням факторного аналізу. Основні поняття факторного аналізу: фактор, факторне навантаження, факторне значення, загальність та специфічність. | <i>Лекція / аудиторна</i> | Презентація лекції | опрацювати лекцію, переглянути презентацію, 2 год |
| | СРС. Оцінювання параметрів факторних моделей. | <i>СРС</i> | Конспект лекції | опрацювати лекцію, 4 год |
| | ПЗ 11. Зниження розмірності даних методом головних компонент. | Практичне заняття | Конспект лекції, завдання на ПЗ 11, | опрацювати лекцію, ознайомитись з літературою, 2 год. |

| | | | | |
|--------------------------------|---|----------------------------------|---|---|
| | | | Field, Andy (2009). Discovering Statistics Using SPSS. – 855 p., Field, Andy (2012). Discovering Statistics Using R. – 995 p. | |
| Тиж. 12 8 год | <i>Лекція 12.</i> Порівняльний аналіз різних методів факторного аналізу. Розвідувальний та конфірматорний факторний аналіз. Факторний аналіз якісних (категоризованих) даних. | Лекція <i>Аудит.</i> | Презентація лекції | опрацювати лекцію, переглянути презентацію, 2 год |
| | СРС. Інтерпретація факторних навантажень. Оцінювання загальності. | <i>СРС</i> | Конспект лекції | опрацювати лекцію, 4 год |
| | Контрольна робота | <i>Контр. робота / аудиторна</i> | Варіанти завдань КР | Повторити матеріал розділів 1-2 |

****якщо література подається в скороченому вигляді, то розшифрування подає в кінці*

11. Система оцінювання та вимоги
Загальна система оцінювання курсу.

| Поточний контроль, самостійна робота, індивідуальні завдання | | | | | | | Контрольна робота, передбачена навчальним планом | Індивідуальні завдання | Разом | Залікова робота | Сума |
|--|----|----|----|----------|----|----|--|------------------------|-------|-----------------|------|
| Розділ 1 | | | | Розділ 2 | | | | | | | |
| T1 | T2 | T3 | T4 | T5 | T6 | T7 | | | | | |
| 5 | 10 | 10 | 5 | 10 | 5 | 5 | 10 | | 60 | 40 | 100 |

T1, T2 ... – теми розділів.

| | |
|---|--|
| Загальна система оцінювання курсу | <i>участь в роботі впродовж семестру/залік - 60/40</i> Розподіл балів, що присвоюються студентам з навчальної дисципліни, є сумою балів за виконання всіх видів завдань та самостійну роботу плюс бали, отримані під час екзамену. Впродовж семестру студент за виконання всіх завдань може отримати до 60 балів і до 40 балів під час складання заліку. |
| Вимоги до контрольної роботи | Завдання з контрольної роботи повинні бути виконані в повному обсязі. При оцінюванні роботи враховується повнота виконання, здатність студента формулювати закони та закономірності, структурувати судження, обґрунтовувати виконані операції, алгоритми, правила визначення понять, робити висновки, описи, тощо. Максимальна оцінка за контрольну роботу – 10 балів. Студент може скласти залік лише за наявності виконаних практичних завдань. |
| Умови допуску до підсумкового контролю | Умовою допуску студента до підсумкового семестрового контролю (екзамену) є наявність усіх виконаних практичних робіт та контрольної роботи. Згідно рішення кафедри теоретичної та прикладної системотехніки факультету комп'ютерних наук до екзамену не допускаються студенти, які мають заборгованість контрольним роботам. Наприкінці курсу всі бали підсумовуються і студент має можливість (в разі успішного виконання всіх завдань) отримати максимальну підсумкову оцінку 100 балів. |

Шкала оцінювання

| Сума балів за всі види навчальної діяльності протягом семестру | Оцінка | |
|--|---------------|---------------|
| | для екзамену | для заліку |
| 90 – 100 | відмінно | зараховано |
| 70-89 | добре | |
| 50-69 | задовільно | |
| 1-49 | не задовільно | не зараховано |

12. Рекомендована література

Основна література

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрии. – М.: ЮНИТИ, 1998. – 1000 с.
2. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. – К.: МОРИОН, 2002. – 640 с.
3. Field, Andy (2009). Discovering Statistics Using SPSS. – 855 p.
4. Field, Andy (2012). Discovering Statistics Using R. – 995 p.
5. Р. И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил.

Допоміжна література

1. Анализ данных/ Н.С. Бакуменко, О.С. Радивоненко. – Учеб. пособие по лабораторному практикуму.–Харьков: Нац. аэрокосм. ун-т «Харьк. авиац. ин-т.», 2007.– 87 с.
2. Решение экономических задач с использованием статистических пакетов обработки данных: учеб. Пособие по лаб. практикуму / М.С. Мазорчук, Н.С. Бакуменко. – Харьков: Нац. аэрокосм. ун-т «Харьк. авиац. ин-т», 2008. – 95 с.
3. Анализ данных результатов статистических исследований в среде R [Текст] : учеб. пособие / М. С. Мазорчук, В. С. Добряк, Н.С. Бакуменко. – Х.: Нац. аэрокосм. ун-т им. Н.Е. Жуковского «Харьк. авиац. ин-т», 2018. – 36 с.
4. Статистический анализ данных в социологии [Текст] : учеб. пособие / М. С. Мазорчук, В. С. Добряк. – Х. : Нац. аэрокосм. ун-т им. Н.Е. Жуковского «Харьк. авиац. ин-т», 2012. – 36 с.