Този курс е предвиден за студенти и докторанти, които в своите магистърски или докторски тези се сблъскват с потребността да извършат определени експерименти, а след това статистически да обработят получените резултати. Оценка за успешно завършване на курса се получава след полагане на тест.

В съвременния живот нуждата от обработка на информация постоянно нараства. Налага се да бъдат вземани решения в множество ситуации от ежедневието ни. От своя страна, всяко решение е толкова по-успешно, колкото по-информирано е взето то. Статистическата обработка на събраната информация е една от основите за вземането на информирани решения. В областта на статистическата обработка съществуват множество софтуерни решения, като се започне от по-достъпните за хора без опит, като *Microsoft Excel* и се стигне до професионалните пакети, като *SPSS*, *Matlab* и *Mathematica.*

Настоящият курс представя програмния продукт R, който първоначално се разработва от Robert Gentleman и Ross Ihaka в University of Auckland през 1993 година. *R* е замислен като алтернатива на програмния продукт *S*, създаден от John Chambers, служител в Bell Labs. Първоначалният замисъл за *R* е да представлява инструмент, който да се използва в интерактивен режим, през командния ред. В последствие тази идея прераства в самостоятелен програмен език. Основното предназначение на *R* е обработка на данни, което включва въвеждане, пресмятане, визуално представяне на графики и отчети.

Езикът получава значително по-голяма популярност след 2000 година, като излиза от рамките на академичните среди и навлиза във финансовите среди, маркетинга, фармацията, социологията, психологията и в много други области. Най-често потребителите на *R* са хора с опит в програмни езици, като *C/C++*, *Java*, *C#* или пък преди това са използвали други статистически пакети, като *SAS*, *SPSS* или дори *Microsoft Excel*. Тези потребители дават значителен тласък в развитието на пакета *R*, добавяйки множество софтуерни приставки (add-ons).

В някои случаи *R* се оказва стряскащ и дори смущаващ, особено за начинаещите потребители, но с времето и с процеса по навлизане в материята овладяването му се улеснява и ускорява. Целта на това учебно помагало е да представи информацията по един достъпен и олекотен за възприемане начин. Изложени са предимно най-важните аспекти от използването на пакета *R*, което от своя страна дава стабилна основа за бъдещо самостоятелно развитие на читателя. Въпреки наличието на голям избор от

учебни материали по *R*, материалът тук е съобразен основно със съдържанието на курса „Анализ на данни с R“, провеждан в Център за обучение към Българска академия на науките.

**Курсът е базиран на учебното помагало “Статистическа обработка на данни с R. Практическо ръководство”, с автори Тодор Балабанов, Зорница Атанасова и Румен Кетипов**, и е организиран в следните части:

**Инсталация и стартиране**: Представя процеса по изтегляне, инсталиране и стартиране на програмния продукт.

**Пакетна организация, променливи, основни математически операции в R и типове данни**: Разяснява основните концепции за работата с програмния продукт *R*, като акцентите са върху пакетната организация на продукта и как най-ефективно да бъдат използвани възможностите му. Разглеждат се базовите типове данни и основните математически операции.

**Сложни структури от данни и извикване на функции**: Демонстрира използването на функции в *R*. Разглеждат се също сложни типове данни, като вектори, фактори, извадки от данни, масиви и матрици.

**Въвеждане на данни и извеждане на графики**: Представя възможностите на системата за въвеждане на данни от външни източници, чрез прочитане на файлове или ресурси в Глобалната мрежа. Демонстрират се основните възможности на програмния продукт за визуализиране на данни и получени резултати.

**Оператори за контрол на изпълнението и потребителски функции**: Въвежда основни принципи, използвани в конвенционалното програмиране като оператори за преход – *if*, *else* и *switch*, цикли – *for* и *while*, както и структуриране на кода под формата на потребителски функции.

**Групиране и обхождане на данни**: Представя възможностите за прилагане на функции върху данни. Показват се групиране на данните и прилагане на агрегатни функции, като *sum*, *min*, *max* и други. Демонстрират се използването на данни във фреймовете и таблици. Демонстрират се възможностите за директна работа с релационна база данни и за обхождане по редове.

**Реорганизация на данните и обработка на символни низове**: Демонстрират се възможностите за сливане на данни, трансформация на колони в редове и обратното.

Разглеждат се възможностите за обработка на символни низове, което включва конструиране на символни низове, търсене на информация в символни низове и как могат да се използват регулярни изрази.

**Разширени графични възможности и вероятностни разпределения**: Представят се по-разширени възможности за графично оформление на информацията. Демонстрират се най-често използваните вероятностни разпределение. Обсъжда се значението на кумулативните и плътностните функции. Демонстрират се възможностите за анализ на случайна величина. Разглеждат се нормално разпределение, биномно разпределение и Поасоново разпределение.

**Статистическа обработка на данните**: Въвежда в статистическата обработка на информацията. Първоначално представя описателните статистики като средна, медиана и мода. Следва представяне на сравнителните статистики като F-тест, T-тест и ANOVA. Завършва с линеен регресионен анализ.

**Приближени пресмятания – подходи, методи, алгоритми**: Посветена е на най-популярните подходи, методи и алгоритми за приближени числени пресмятания. С помощта на Монте-Карло методите се извършват множество статистически оценки в различни сфери от икономиката. При глобална оптимизация в многомерни пространства генетичните алгоритми са един от най-ефективните инструменти за търсене на субоптимални решения, близки до глобалните оптимуми. При необходимост от обработка на информация, спрямо съществуващи примерни, но без ясно формулирани правила за пресмятане, изкуствените невронни мрежи показват задоволителни резултати.

**Оформление на резултатите за печатно и електронно представяне**: Представени са възможностите за оформление на получените резултати като предпечатна подготовка и представяне с мултимедийна техника. В основата на визуалното оформление с програмния продукт R са заложени тагиращите езици *LaTeX* и *Markdown*. В резултат на финалната компилация продуктите могат да бъдат PDF файлове, HTML страници, слайдове за презентация и текстови документи.

Разгледаните теми в учебното помагало дават възможност на заинтересованите потребители да започнат като напълно начинаещи работа с програмния продукт *R*. Макар и полезни, предварителни знания по програмиране не са необходими, но е желателно да са налични предварителни знания по статистика. Изложеният материал превежда читателите от темите за начинаещи до потребители в средно ниво. Учебното

помагало не засяга темите за напреднали, които включват сложните методи за статистическа обработка, като нелинейни регресионни модели и дървовидни структури на решенията. Също така, не се засягат темите за напреднали по отношение на самия програмен продукт *R* и неговите възможности като самостоятелен програмен език, какъвто пример е темата за програмирането и добавянето на нови пакети (софтуерни библиотеки).

Авторите изразяват своята надежда, че всеки заинтересован читател ще намери полезна информация в настоящото учебно помагало, която ще му позволи да разшири своите знания, умения, а и цялостен мироглед към света и живота. **Авторите изказват най-искрени благодарности към колегите доц. д-р Вера Ангелова и Нина Керемедчиева, за безценното съдействие, което указаха в процеса по създаването на това учебно помагало.**