Лабораторна робота 1.

Вам необхідно здійснити обробку вибіркових даних в наведених нижче завданнях. Обробку необхідно здійснювати в декілька етапів (вони відповідають різним темам курсу). Виконання ЛР1 відповідає першому етапу. На цьому етапі ви маєте практично закріпити знання отримані в рамках теми «Описова статиста в пакеті R».

В усіх наведених нижче завданнях, для заданих вибірок, вам необхідно виконати такі основні ***практичні задачі***:

1. обчислити відомі вам міри центральної тенденції та міри мінливості (середнє значення, медіана, стандартне відхилення, розмах, першу та другу квартилі, міжквартильний інтервал), виведіть їх на екран разом із текстом, що прояснює, що за значення на екрані (команда cat)
2. побудувати гістограму. Гістограма має бути анотована (заголовок – вказати, що за данні, підпис осі x та y – вказати, що за величини). Якщо вибірок декілька – їх гістограми мають бути розташовані одна під одною, мати однаковий масштаб по осі абсцис.
3. побудувати коробчасту діаграму («ящик з вусами»). Якщо вибірок декілька – всі вибірки мають бути представлені на одній діаграмі. Діаграма має бути анотована (заголовок – вказати, що за данні порівнюються; підпис осі x – вказати що за вибірка, та y – вказати, що за величина).

Кожне завдання робіть в окремому записнику. Всі завдання ви можете виконувати використовуючи записник для першого завдання як шаблон (скопіюйте його в новий записник, поміняйте набір даних, підписи на графіках і т.д.)

Теоретичні питання:

1. Використовуючи гістограму сказати - який розподіл має кожна з випадкових величин (симетричний, несиметричний чи є викиди)
2. Порівняйте вибіркові середні та медіани, розсіяність вибірок. Виходячи з форми розподілу, обґрунтуйте який з показників краще використовувати в якості міри центральної тенденції (середнє чи медіану) та розсіяності (стандартне відхилення, міжквартильний інтервал)
3. Вміти оцінити з гістограми приблизні значення: середнього, медіани, стандартного відхилення, розмаху, міжквартильного інтервалу
4. Вміти оцінити з коробчастої діаграми приблизні значення: медіани, міжквартильного інтервалу, наявність викидів
5. Для заданої вибірки вміти вручну обчислити: медіану, міжквартильний інтервал.
6. Використовуючи коробчасту діаграму перенесіть результати порівняння вибірок на генеральну сукупність (оцініть, чи можна результати порівняння вибірок перенести на генеральну сукупність?).

Завдання 1. Забрудненість озер

В трьох озерах було відібрано по 20 проб води, в кожній пробі було виміряно рівень забруднення води певною речовиною. Рівень забруднення для всіх проб наведено в таблиці:

Вибірка	Рівень забруднення, мг/л
Озеро 1	2.7085, 4.5291, 1.9241, 4.8117, 0.69854, 3.2057, 3.5454, 3.4412, 3.2103, 2.5617, 3.9917, 1.8849, 3.2781, 4.1563, 2.0624, 3.9661, 1.902, 2.0625, 1.6876, 3.6103
Озеро 2	2.4822, 4.4128, 3.7624, 4.254, 3.0504, 2.5828, 2.3083, 3.3501, 4.7932, 3.3775, 3.1039, 3.9196, 3.5303, 3.2286, 4.8949, 4.9019, 4.6743, 2.5583, 3.6254, 2.9874
Озеро 3	8.9456, 7.0399, 7.1279, 6.6328, 9.8243, 5.6612, 6.8074, 6.1613, 10.183, 7.6435, 6.0573, 8.6647, 8.4496, 8.204, 6.7038, 7.9422, 6.9381, 8.2265, 9.092, 9.9832

1. Вибірки з якого озера є найбільш та найменш забрудненими?
2. Яке озеро є найбільш та найменш забрудненим?
3. В чому різниця між двома попередніми запитаннями?

4. Вибірки з якого озера мають найбільшу та найменшу розсіяність, що це означає?
5. Чому необхідно обчислювати розсіяність – яким чином це може допомогти дати відповідь на питання 2

Завдання 2. Чи знання це сила?

На металургійному заводі «Світлий шлях» були проведені дослідження зв'язку між оцінками що робітники отримують при здачі правил техніки безпеки та наявністю травм що були отримані на виробництві. Нижче наведені результати підрахунку сумарної кількості пальців на обох руках у слюсарів що мали оцінку менше 40 балів (за 100 бальною шкалою):

1,1,2,1,3,5,4,5,1,7,6,5,4,1,5,3,5,9
В цю групу потрапило 18 робітників.

Також було отримана сумарної кількості пальців на обох руках у слюсарів що мали оцінку більше 40 балів:

9,8,3,10,10,8,10,15,10,9,9,10,10,10,8,1,4,5,3,8,9,9,10,10,10,7,7,6,8,10,9,8

Порівняйте вибірки між собою.

Чи можна сказати, що рівень оцінки на екзамені по техніці безпеки впливає на кількість пальців у людини.

Завдання 3. Завдання 3. Скільки кави п'ють люди.

В Китаї та Бразилії було опитано по 100 чоловік, на предмет того, скільки зазвичай за день вони випивають чашок кави. Були отримані наступні данні:

Китай:

3,2,1,3,7,4,5,5,2,5,0,7,5,1,0,6,4,10,4,5,2,2,4,5,1,2,5,2,1,0,5,2,14,0,3,8,1,2,1,2,8,4,1,5,1,1,3,6,1,5,2,5,2,5,1,4,
1,2,2,7,7,9,3,5,5,0,5,0,2,11,6,6,23,0,2,1,1,4,2,1,0,8,3,1,9,3,0,0,0,6,8,1,4,3,3,3,1,3,2,3

Бразилія:

7,6,0,7,11,1,9,9,6,9,3,11,9,5,1,10,8,14,8,9,6,6,8,4,1,3,4,1,5,4,9,6,3,1,7,12,5,6,5,6,3,8,5,9,0,5,7,10,5,9,6,4,3,9
,5,8,5,6,6,3,11,13,7,9,9,3,9,3,6,15,10,10,27,4,6,0,5,8,6,5,4,12,1,5,13,7,3,3,4,10,12,3,8,7,7,7,5,7,6,7

Порівняйте вибірки між собою.

Чи можна сказати, що китайці чи бразильці п'ють більше кави?

В послідуючих завданнях вибіркові данні знаходяться в текстових файлах або в файлах Excel. Вам буде необхідно завантажити ці данні з цих файлів в пам'ять програми R. Завантажену таблицю потрібно вміти використовувати. Як це все робити можна подивитись у відео **ЛР1 Відео Доданок. Фрейми даних**. або прочитати в документі **ЛР1 ДоданкиФайл: Доданок. Експорт даних; Доданок. Як з таблиці отримати дані тільки для певної категорії**.).

Самі файли з даними можна скачати як архів. Якщо ви працюєте в хмарі rstudio.cloude закачайте **ЛР1 Файли (rstudio.cloude)** інакше **ЛР1 Файли (RStudio Desktop)**

Завдання 4. Чи впливає положення тіла на кров'яний тиск

Проводився експеримент в якому з'ясували вплив положення тіла на кров'яний тиск. У 32х учасників вимірювали кров'яний тиск. Тиск вимірювали в двох положеннях:

1) піддослідний лежав горілиць, руки лежали вздовж тіла
2) людина вставала та тримала руки на рівні серця

Результати вимірювань наведені в таблиці в файлі blood_pres.txt. Файл blood_pres.txt має такі стовбці:

«**піддослідний**» - ПІБ піддослідного

«**лежачі_с**» та «**лежачі_д**» - систолічний та діастолічний тиск виміряний в лежачому положенні

«**стоячі_с**» та «**стоячі_д**» - систолічний та діастолічний тиск виміряний стоячі (тиск в мм. рт. ст.).

Не обчислюйте статистичні характеристики (середнє, медіана, ….) окремо для двох наборів значень **систолічний тиск лежачи**, **систолічний тиск стоячи**. Замість цього обчисліть їх для різниці: **систолічний тиск лежачи – систолічний тиск стоячи**. Цеж стосується гістограми та ящика з вусами (на ящику з вусами буде тільки один «ящик»). Поясніть чому це так. Як інтерпретувати ящик з вусами, як оцінити достовірність висновку? Так само зробіть з наборами значень **діастолічний тиск лежачи**, **діастолічний тиск стоячи**.

Чи залежить тиск (окремо систолічний та діастолічний) від положення тіла?

Завдання 5. Завдання 5. Титанік.

Судно «Титанік» затонуло в 1912 році, більшість пасажирів загинуло. Було отримана детальна інформація стосовно 1309 пасажирів та екіпажу на борту судна Титанік. Ці данні були використанні в чисельних статтях в пресі, включаючи цю:

More Britons than Americans died on Titanic 'because they queued' (На Титаніку загинуло більше британців ніж американців, оскільки перші «стояли в черзі») https://www.independent.co.uk/news/world/australasia/more-britons-than-americans-died-on-titanic-because-they-queued-1452299.html

Ця стаття приклад невірної інтерпретації статистичних результатів. Фактор національності був важливим для виживання тільки відповідно до критерію хі квадрат, побудова ж логістичної регресійної моделі показала що більш значущім фактором був клас в якому подорожував пасажир. Більшість американців подорожувала першим класом.

Данні представлені в таблиці в файлі Titanic.csv. В стовпчиках таблиці знаходяться таки величини:

Назва стовпчику	Пояснення	Тип даних	Можливі значення величини
Pclass	Клас	Ordinal	1 = 1й, 2= 2й, 3 = 3й
Survived	Чи вижив пасажир	Binary (Nominal)	0 = Помер, 1 = Вижив
Residence	Громадянство	Nominal	0 = American, 1 = British, 3 = Other
Name		String	
Age	Вік	Scale	
Sibsp	Кількість братів і сестер / подружжя	Scale (Discrete)	
Parch	Кількість батьків / дітей на борту	Scale (discrete)	
Ticket	Номер квитка	String	
Fare	Ціна квитка	Scale	
Cabin	Номер кабіни	String	
Embarked	Where passenger embarked	String	
Boat	Boat identification (if rescued)	String	
Body	Номер тіла (якщо помер)	ID	
Home.dest	Home town	tring	
Gender	Стать	Binary (Nominal)	0 = чоловік, 1 = жінка

Вибірки які ви будете використовувати в цій роботі можуть мати значення «NA» - не число. Це означає, що в таблиці з даними для певних пасажирів відсутні данні. Для таких вибірок, функція mean не зможе порахувати середнє, це стосується і деяких інших функцій. Як цьому запобігти: читайте **Доданок. Що робити якщо відсутня частина даних?** або дивіться **ЛР1 Відео Доданок. Експортданих**

Чи впливала на смертність ціна квитка пасажира? (буде дві вибірки: 1) ціни квитків пасажирів, що вижили та 2) ціни квитків пасажирів, що загинули)

Чи впливав вік на смертність?

Як захищати індивідуальні завдання

На захист 2х завдань максимум 15 хв на людину. Про результати слід розказувати строго дотримуючись наведеного нижче плану. Паралельно все, що ви розказуєте ви ілюструєте результатами з вашого записника (всі результати мають бути наявними, перевірте це до захисту). Для того, що би вкластись у відведений час, доповідь ви маєте підготувати заздалегідь вдома, передавившись роботу, обдумавши, що ви скажете по кожному з пунктів.

1) Коротко перекажіть завдання, які саме задачі вам потрібно вирішити
2) Охарактеризуйте вибірки: a) що це за вибірки (чого ми наобирали); b) об’єм; c) тип розподілу (висновок про тип розподілу підтвердити всіма доступними методами)
3) Зробіть попередні статистичні висновки (в усіх випадках вам необхідно не просто оголосити висновок, але і аргументувати як він витікає з отриманих результатів) використовуючи:

   a) вибіркові статистичні показники

   b) гістограми

   c) коробчасті діаграми
4) Якщо це можливо, зробіть статистичні висновки використовуючи довірчі інтервали (ДІ). Якщо ви розрахували (чи не розрахували) ДІ – докажіть, що він може бути (не може бути) обчислений для вашої вибірки (буде чи ні приблизно виконуватись центральна гранична теорема?). Скажіть що показує довірчий інтервал.

5) Зробіть статистичні висновки використовуючи статистичні тести. Обґрунтуйте вибір тесту. Вам необхідно не просто оголосити висновок, але і аргументувати як він витікає з отриманих результатів, сказати якою є його достовірність.

Завдання 6. Вага новонароджених (чи корисно палити)

Файл birthweight.csv містить набір даних що стосується інформації про новонароджених дітей та їхніх батьків. Він містить в основному безперервні змінні (хоча деякі з них мають лише кілька значень, наприклад, кількість сигарет що випалюються за день.

Основна залежна величина = Birthweight (фунти)

Name	Variable	Data type
ID	Номер дитини	
length	Висота дитини (дюйми)	Scale
Birthweight	Вага дитини (фунти)	Scale
headcirumference	Окружність голови	Scale
Gestation	Період вагітності (тижні)	Scale
smoker	Мати палить = 1, мати не палить = 0	Binary
motherage	Вік матері	Scale
mnocig	Кількість цигарок що їх випалює мати на день	Scale
mheight	Висота матері (дюйми)	Scale
mppwt	Вага матері до вагітності (фунти)	Scale
fage	Вік батька	Scale
fedyrs	Father's years in education Тривалість часу (роки) що батько витратив на освіту	Scale
fnocig	Кількість цигарок що їх випалює батько на день	Scale
fheight	Висота батька (дюйми)	Scale
lowbwt	Мала вага при народженні, 0 = Ні та 1 = Так	Binary
mage35	Вік матері за 35 років, 0 = Ні та 1 = Так	Binary

Побудуйте діаграми кількості матерів що палять.

Основні задачі вирішить окремо для двох наборів даних: 1) зі стовпчика Birthweight візьміть тільки ті значення які стосуються дітей матері яких не палили. 2) зі стовпчика Birthweight візьміть тільки ті значення які стосуються дітей матері яких палили. (див. ***Доданок. Як з таблиці отримати дані тільки для певної категорії***). Для двох гістограм зробіть однакові границі вздовж осі ікс, див. ***Доданок. Гістограма як задати масштаб осей***. Чи впливає те чи курить мати на вагу новонародженої дитини?

Побудуйте точкові діаграми (***Доданок. Точкові діаграми***), залежності ваги дитини від періоду вагітності, для дітей матері яких палили та не палили (для цих двох випадків розфарбуйте точки двома різними кольорами) . Анотуйте графік. Обидві діаграми мають бути побудовані на одній координатній осі. Інтерпретуйте отриманий графік . Яким чином слід враховувати залежність ваги малюка від тривалості вагітності при вирішенні того чи впливає паління на вагу?

Завдання 7. Завдання 4. Скільки кави п’ють люди (варіант 2)

Як виконати це завдання читайте в ***Доданок. Імпорт даних***

Виконайте попереднє завдання завантаживши вибіркові данні з таблиці з файлу **Cofee.csv**. Таблиця в цьому файлі має два стовпчика:

N	**кільк_чашок**
1	3
2	2
3	1
…	…

N – номер людини що приймала участь в опитуванні

кільк_чашок – скільки чашок на день вона випиває (нам потрібен буде тільки цей стовпчик)

Зауваження:

- Файл **Cofee.csv** знаходиться в середині архіву **lab1_data.zip** (разом з іншими файлами які ви будете використовувати далі в цій лабораторній). Скачайте цей файл собі на локальний диск, після чого завантажте його в хмару, так як ви це робили з файлом diet_examp.txt в ***Доданок. Імпорт даних***. Коли ви завантажите цей архів в хмару, він автоматично буде розархівований (перевірте на закладці **Files** що в списку файлів з’явився файл **Cofee.csv**. Відкрийте його, подивіться вміст, закрийте.
- До записника з цим завданням включіть команду **library('rio')**
- Імпортуйте таблицю з файлу **Cofee.csv** за допомогою команди **tabl<-import('Cofee.csv')**
- В попередньому завданні вибіркові данні ви зберігали в векторі **d**. Тепер ті самі данні будуть знаходитись в стовпчику **кільк_чашок** фрейму даних **tabl**. Використовуючи фрейм даних **tabl**, зробіть з цими даними те саме що ви робили з ними в попередньому завданні (гістограма, числові показники мають бути такими ж як і в попередньому завданні).

Завдання 8. Завдання 7. Досліджуємо коріння помідорів

Експериментатор цікавиться генетикою процесу формування кореневої системи томатних рослин. Він працював з популяціями практично ізогенних ліній (тобто майже генетично ідентичних), які були отримані шляхом схрещування між інбредним комерційним сортом (M82) та дикою лінією. Ці лінії генетично ідентичні M82, за винятком одної ділянки (різний для кожної лінії), де лінії мають гени дикого родича. Він виділив дві лінії (A та B), які, здається, відрізняються від M82 в деяких властивостях вкорінення. Експериментатор виконував два експерименти, щоб більш детально вивчити властивості формування кореневої системи цих двох ліній відносно M82, а також зв'язок між розвитком коріння та пагонів.

а) У першому експерименті вивчалась здатність коріння проникати через бар'єр. Експериментатор вирощує помідори в окремих пластмасових трубках з компостом, 30 см заввишки, з мембраною на дні трубки, яка заважає росту коріння. Коли вік рослини складає декілька тижнів, для кожної рослини дослідник підраховує коріння, яке проникло в мембрану, фіксує довжину, на яку найдовший корінь проникає в мембрану, а потім зрізає коріння безпосередньо під мембраною і зважує їх. Він також зрізає над землею пагіння рослини (від рівня ґрунту) і зважує його. Він висушує в печі зрізане коріння та пагони, а потім знову зважує висушений матеріал

b) У другому експерименті рослини вирощують гідропонічно (не в грунті), в окремих високих (2м довжиною) трубках. Коли рослини починають квітнути, дослідник видаляє їх з трубок і вимірює довжину найдовшого кореня. Потім він розрізає коріння на чотири секції, кожна з яких має одну чверть довжини найдовшого кореня, так що перша секція - це чверть коренів, найближчих до поверхні, а четверта ділянка - це чверть, що знаходиться якнайдалі від поверхні. Ці секції сушать і зважують. Пагони також зважують до і після сушіння.

Результати першого експерименту представлені в таблиці в файлі Rdataset-Tomato1.csv. Ця таблиця має наступні стовпчики:

Стовпчик	Що в стовпчику		Тип даних
Genotype	Line (variety)	Лінія (сорт)	Категоріальна
NORP	Number of penetrating roots	Число проникаючих коренів	дискретна
RootFwt	Penetrating root fresh weight (g)	Проникаючі корні свіжа вага (г)	Безперервна
RootDwt	Penetrating root dry weight (g)	Проникаючи корні, суха вага (г)	Неперервна
CanoptFwt	Canopy fresh weight (g)	Пагони свіжа вага (г)	Безперервна
CanopyDwt	Canopy dry weight (g)	Вага сухих пагонів (г)	Безперервна
Length	Root length below the membrane (mm)	Довжина кореня нижче мембрани (мм)	Безперервна

Результати першого експерименту представлені в таблиці в файлі Rdataset-Tomato2.csv. Ця таблиця має наступні стовпчики:

Стовпчик	Що в стовпчику		Тип даних
Tubenumber			
Genotype	Line (variety)	Лінія (сорт)	Категоріальна
Finalrootlengthmm	Final root length (mm)	Остаточна довжина кореня (мм)	безперервна
RootdwtgSection1	Root dry weight (section 1) (g)	Суха вага коріння (ділянка 1) (г)	Безперервна
RootdwtgSection2	Root dry weight (section 2) (g)	Суха вага коріння (ділянка 2) (г)	Безперервна
RootdwtgSection3	Root dry weight (section 3) (g)	Суха вага коріння (ділянка 3) (г)	Безперервна
RootdwtgSection4	Root dry weight (section 4) (g)	Суха вага коріння (ділянка 4) (г)	Безперервна
Canopyfwtg	Canopy fresh weight (g)	Вага свіжих пагонів (г)	Безперервна
Canopydwtg	Canopy dry weight (g)	Вага сухих пагонів (г)	Безперервна
TotalRootdwtg	Total root dry weight (g)	Сумарна вага сухого коріння(г)	безперервний

В рамках вирішення основних завдань покажіть чи відрізняється середня довжина коріння для трьох ліній ? Чи відрізняється проникаюча здатність коренів для трьох ліній ? Обґрунтуйте вибір величин що ви обрали для порівняння.

There are a number of possible response variables, though main focus is the effect of line on rooting properties, and the effect of rooting properties on canopy development.

Can be used for:

Method/Test	Questions
Descriptive statistics	What is the mean root length for each line? What is the variability in penetrating root dry weight for each line?
Box-and-whisker plots	Compare the distributions of penetrating root dry weights between the lines
Confidence	Construct an interval within which the true (population) mean total root dry

intervals	weight for M82 would fall (with 95% confidence)
Normal probability calculations	Assuming that canopy dry weight is Normally distributed, calculate the proportion of plants in the population with a canopy dry weight greater than x g, or is between x and y g
Correlation coefficients	Is there a linear association between canopy dry weight and root dry weight? Is there a linear association between root fresh weight and root dry weight? Is there a linear association between penetrating root dry weight and number of penetrating roots?
Scatter plots	Is there an association between canopy dry weight and root dry weight? Is there an association between root fresh weight and root dry weight? Is there an association between penetrating root dry weight and number of penetrating roots?
One-sample t-test	Are the roots of N82 grown hydroponically (second experiment) of a similar length to what we would expect for plants grown in soil?
Two-sample t-test	Are there differences in the mean root lengths between lines A and B? (could include both the equal variance case and the unequal variance (Welch test) case
F-test	Are there differences in the variances of root lengths between lines And B? (also directly associated with the assumption of equal variance for the two-sample t-test)
One-way ANOVA	Are there differences between the mean root lengths for the three lines (M82, A, B)?
Simple linear regression	Is the size of the canopy (canopy dry weight) linearly related to the amount of roots (total root dry weight)? Is the penetrating root dry weight linearly related to the number of penetrating roots? Is the root fresh weight linearly related to the root dry weight?
Multiple linear regression	Can the prediction of the size of the canopy be improved by considering the weights of roots at different depths below the soil surface?
Simple linear regression with groups	Does the relationship between canopy dry weight and root dry weight vary between lines? Does the relationship between root fresh weight and root dry weight vary with line? Does the relationship between penetrating root dry weight and number of penetrating roots vary with line?
Multiple linear regression with groups	Does the relationship between canopy dry weight and the weights of roots at different depths below the soli surface vary with line?