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

Дайте відповідь на питання чи впливає рівень оцінки на екзамені по техніці безпеки на кількість пальців у людини.

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

Було опитано 100 людей, на предмет того, скільки зазвичай за день вони випивають чашок кави. Були отримані наступні данні:

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

Дайте відповідь на питання, скільки в середньому на день, люди вживають чашок кофе? Обґрунтуйте, які оцінки показників центральної тенденції краще використовувати для відповіді на це питання (середнє арифметичне або медіану).

З надійних джерел відомо: в день люди в середньому випивають 2 чашки кави. Чи підтверджує це твердження експеримент наведений в завданні? (не забувайте, що вибіркове середнє чи медіана самі по собі кажуть тільки про людей вибірки, не кажуть взагалі про всіх людей)

З іще більш надійних джерел відомо: в день люди в середньому випивають 6 чашок кави. Чи підтверджує це твердження експеримент наведений в завданні?

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.