Лабораторная работа № 1. Первичный анализ данных
Срок сдачи до 07.10.2021.

По результатам выполнения лабораторной работы необходимо подготовить отчет в электронном виде. В качестве отчета вполне подойдет код на языке R или Python с комментариями и результатами вычислений.

Основные задания **(на 4 балла)**

0. Загрузить выборку из файла “VarN.txt”, где N – номер варианта, он же номер в списке подгруппы (см. таблицу с текущим рейтингом).
1. Провести визуальный анализ выборки, то есть вывести: а) график реализации, б) ящик с усами, в) гистограмму (постройте несколько гистограмм с разным числом интервалов: а) небольшим: 2-3; б) средним – обычно выбирается автоматически, в) большим: порядка 30 или даже более). Проинтерпретировать полученные иллюстрации. Попробуйте предположить, к какому распределению принадлежит выборка.
2. Вычислить числовые характеристики выборки: среднее, стандартное отклонение, дисперсию, медиану, первую и третью квартиль, коэффициент асимметрии, эксцесса, минимальное и максимальное значение выборки. Какие можно сделать выводы?

Бонусные задания

Проверка правила 3σ.

3. **(1 балл)** Вычислите, какова доля тех наблюдений, которые попали в интервал $[\bar{x} - 3s, \bar{x} + 3s]$, где $\bar{x}$ – выборочное среднее, s – выборочная оценка стандартного отклонения. Что можно сказать про теоретическую вероятность $P\{\xi \in [\mu - 3\sigma, \mu + 3\sigma]\}$. (См. Неравенство Чебышева). Согласуется ли практический результат с практикой?

Подгонка закона распределения

4. **(1 балл)** По гистограмме и вычисленным числовых характеристикам попробуйте сделать вывод о принадлежности выборки к одному из следующих распределений: нормальное, экспоненциальное, Лапласа, Вейбулла, равномерное.
5. **(1,5 балл за тест)** Проверить гипотезы согласия о принадлежности выборки к одному из 5 указанных в пункте 4 распределений. Рекомендуемые тесты: хи-квадрат Пирсона, Колмогорова-Смирнова (можно использовать другие тесты согласия). Сделать вывод о том, к какому распределению она принадлежит, с какими параметрами.

Во всех последующих пунктах, где требуется вычислить некоторые статистики, приветствуется визуализация данных графиками из пункта 1. Однако в этих пунктах достаточно будет обойтись одной гистограммой, а не несколькими, как это требуется в постановке пункта 1.

Работа с пропусками

6. **(0,5 балла)** Случайным образом внести в выборку пропуски (долю пропущенных наблюдений выберете в пределах 0,2 0,5).
7. **(1 балл)** Вычислить числовые характеристики из пункта 2, игнорируя пропущенные значения, т.е. считайте, что у вас новая выборка, состоящая из элементов, которые остались непропущенными. Насколько сильно посчитанные числовые характеристики отличаются от результатов, полученных по полным данным? Почему?
8. **(1 балл).** Заполнить пропущенные значения средним арифметическим по присутствующим данным. Вычислить числовые характеристики из пункта 2. Насколько сильно посчитанные числовые характеристики отличаются от результатов, полученных в пунктах 2 и 7? Почему?
9. **(0,5 балла).** Как надо модифицировать формулы вычисления оценок параметров в случае заполнения пропусков из пункта 8, чтобы они стали близки к соответствующим значениям в случае полных данных?

Цензурирование данных снизу. Цензурирование – механизм порождения пропусков, при котором, сам факт, наблюдается ли текущее значение или нет, зависит от самого наблюдения. Например, в случае цензурирования снизу все значения меньшие некоторого заданного уровня пропускаются.

10. **(0,5 балла)** Выбрать уровень цензурирования $c = (1/3)x_{max} + (2/3)x_{min}$. Все значения, меньшие c заменить на пропуски.
11. **(1 балл)** Вычислить числовые характеристики из пункта 2, игнорируя пропущенные значения. Насколько сильно посчитанные числовые характеристики отличаются от результатов, полученных по полным данным? Почему?
12. **(1 балл).** Заполнить пропущенные значения средним арифметическим по присутствующим данным. Вычислить числовые характеристики из пункта 2. Адекватными ли получаются результаты?
13. **(***)** Как надо модифицировать формулы вычисления оценок параметров в случае заполнения пропусков из пункта 12, чтобы они стали близки к соответствующим значениям в случае полных данных?

Выбросы. Выброс (outlier) — в статистике результат измерения, выделяющийся из общей выборки.

14. **(1 балл)** Пусть $L = x_{min}\ \ x_{max}$, тогда добавить в выборку 5 элементов, больших или равных $x_{max} + L$, и 5 элементов, меньших или равных $x_{min}\ \ L$. Вычислить числовые характеристики из пункта 2. Насколько сильно посчитанные числовые характеристики отличаются от результатов, полученных по исходным данным? Почему?
15. **(1 балл)** Добавить в выборку 10 элементов, больших или равных x_{max} + L. Вычислить числовые характеристики из пункта 2. Насколько сильно посчитанные числовые характеристики отличаются от результатов, полученных по исходным данным? Почему?