Гайд по ЛР2 (матстат)

Задание 1

Необходимая база

Для выполнения сей лабораторной очень стоит прочитать про такие вещи, как распределение χ^2 (хи-квадрат), распределение Стьюдента и распределение Фишера на википедии, ну или здесь в моих кратких и не всегда формальных формулировках.

Распределение χ^2

Пусть есть совокупность случайных независимых стандартных нормальных величин $X_1, \dots, X_n \sim N(0,1)$.
Тогда сумма квадратов этих величин будет иметь распределение хи-квадрат с n степенями свободы:

$$X = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

Степени свободы == количество слагаемых/квадратов, которые могут свободно изменяться в контексте вычисления величины какого-либо параметра.

Например, величина $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$, т.е. имеет распределение хи-квадрат с $n-1$ степенью свободы.

Распределение Стьюдента t

Пусть $X_0, \dots, X_n \sim N(0,1)$. Тогда величина

$$t = \frac{X_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}} = \frac{X_0}{\frac{\chi_{n-1}^2}{\sqrt{n}}}$$

имеет распределение Стьюдента с $n-1$ степенью свободы, или же $\sim t_{n-1}$.

Распределение Фишера F

Пусть $X_1, \dots, X_n \sim N(0,1),\ Y_1, \dots, Y_m \sim N(0,1)$
Тогда величины $Z_1 = \sum_{i=1}^{n} X_i^2$ и $Z_2 = \sum_{i=1}^{m} Y_i^2$ будут иметь распределения χ_n^2 и χ_m^2 соответственно, а величина

$$F = \frac{\frac{Z_1}{n}}{\frac{Z_2}{m}}$$

будет иметь распределение Фишера со степенями свободы n, m.
Простыми словами: распределение Фишера — это отношение двух величин с распределением хи-квадрат, деленных на свои количества степеней свободы.

Доверительный интервал уровня доверия $1 - \alpha$

Пусть $0 < \alpha < 1$, а $\theta_l(X, \alpha), \theta_r(X, \alpha)$ — некие функции. Тогда интервал (θ_l, θ_r) будет называться доверительным интервалом уровня доверия $1 - \alpha$ для параметра θ, если

$$\forall \theta \in \Theta : P_\theta(\theta_l < \theta < \theta_r) \geq 1 - \alpha$$

Нахрена все это нужно

В первом задании необходимо будет составить доверительный интервал (далее ДИ) для какой-то величины, будь то разность матожиданий или отношение дисперсий. Для составления ДИ для величины нам нужно будет воспользоваться оценками тех параметров, которые участвуют в ее формуле (например, оценкой матожидания, т.е. средним выборочным), а именно проверить, попадает ли составленная оценка для искомой величины в интервал, полученный какими-то математическими преобразованиями, а также с помощью квантилей соответствующих распределений. Именно для этого важно научиться различать все эти распределения и понимать, с каким именно распределением вы имеется дело.
Формулы в вариантах 1-2 очень напоминают распределение Стьюдента, в то время как формулы в вариантах 3-4 явно похожи на распределение Фишера. Определение числа степеней свободы я оставляю на вас, благо это не так сложно.

Собственно, задание

Суть задания

Вам предстоит сгенерировать $2 \cdot 2 \cdot (1 + 1000) = 4004$ выборки и посчитать для них $2 \cdot (1 + 1000)$ ДИ, после чего проверить попадание истинного значения параметра в этот ДИ. Но это потом.

Формула для ДИ

Для начала следует определиться с формулой для ДИ. Во всех вариантах она будет разная и как-то основываться на той подсказке (функции), которая указана в соответствующем варианте.
Соответственно, ДИ для этой функции (статистики) будет заключен между квантилями порядков $\frac{\alpha}{2}$ и $1 - \frac{\alpha}{2}$ соответствующего распределения (Стьюдента или Фишера). Они

ищутся по таблицам, которые легко можно найти в интернете.
Останется только выразить из получившегося двойного неравенства необходимую величину.
Итого, универсальный рецепт для формулы таков:

1. Выбрать функцию, отвечающую какому-то распределению
2. Найти выборочные квантили, соответствующие $\alpha = 0.05$ и другим параметрам выбранного распределения
3. Оценить статистику выборочными квантилями
4. Выразить искомую величину и получить результирующий ДИ
   После этого можно переходить к следующей части задания.

По поводу остальных подсказок

Во всех вариантах есть еще одна серия условий. Например, в третьем варианте:

$$\mu_1, \mu_2 \text{ неизвестны}; \mu_1 = 0, \mu_2 = 0, \sigma_1^2 = 2, \sigma_2^2 = 1$$

На первый взгляд какое-то дерьмо. Матожидания неизвестны, но при этом они известны. Так еще и дисперсия известна, а в этом варианте именно она оценивается. Тогда вообще не надо ничего оценивать? Или как-то по-разному генерировать выборки. На деле все проще. В каждом из вариантов это условие состоит из двух частей — заданных параметров и используемых значений в формуле. Пойдем в обратном порядке:

- $\mu_1 = 0, \mu_2 = 0, \sigma_1^2 = 2, \sigma_2^2 = 1$ — это заданные параметры для генерации распределения. Именно их нужно использовать для генерации выборок в своих программах.
- μ_1, μ_2 неизвестны — данное условие показывает, что именно нужно использовать при вычислении ДИ в программе: истинное значение параметра или же его выборочную оценку.
  - Если указанные величины известны, то используются истинные значения параметров (т.е. те, которые вы использовали при генерации выборки)
  - Если неизвестны, то вы используете выборочные оценки параметров (среднее выборочное для матожидания μ или выборочную дисперсию для дисперсии σ).

Генерация выборок и дальнейшие действия

Дальше все достаточно тривиально:

1. Генерируете 2 выборки X_1 и Y_1 объема 25, после чего вычисляете для нее ДИ.
2. Проверяете, попала ли искомая величина (например, в третьем варианте это $\frac{\sigma_1^2}{\sigma_2^2}$) в ДИ. Для вычисления величины нужно использовать реальные значения параметра

3. Повторяете пункты 1-2 1000 раз и для каждой выборки проверяете, попала ли искомая величина в ДИ. Не забывайте считать количество попаданий. В конце составляете общий процент попадания, т.е. $\frac{\text{кол-во попаданий}}{\text{всего ДИ}} \cdot 100\%$. Должно получиться $\approx 95\%$, т.к. мы брали $\alpha = 0.05$
4. Далее повторяете пункты 1-3 для выборок объема 10000. Также не забудьте посмотреть, как изменится ДИ относительно выборок меньшего объема — по-хорошему, он должен быть меньше, т.к. больше элементов в выборках, следовательно больше точность и уже интервал.

Задание 2

Легкая часть задания

Во втором задании нужно будет построить уже не "обычный" ДИ, а асимптотический (далее АДИ). После этого для полученной формулы АДИ все так же, как в первом задании генерируются выборки, считается процентовка, увеличивается объем и т.д. Для этого достаточно действовать аналогично пунктам в первом задании. Так что в целом задание почти такое же, **но есть один нюанс**.

Сложная часть задания

Асимптотический доверительный интервал

Пусть $0 < \alpha < 1$, а $\theta_l(X, \alpha), \theta_r(X, \alpha)$ — некие функции. Тогда интервал (θ_l, θ_r) будет называться асимптотическим доверительным интервалом (асимптотического) уровня доверия $1 - \alpha$ для параметра $\theta \in \Theta$, если

$$\forall \theta \in \Theta : \lim_{n \to \infty} \inf P_\theta(\theta_l < \theta < \theta_r) \geq 1 - \alpha$$

Простыми словами, АДИ — это такой же доверительный интервал, просто в качестве оцениваемой квантилями функции мы выбираем такую, которая слабо сходится к распределению этих самых квантилей при $n \to \infty$. Соответственно, действительно точным он будет становиться только для больших n.

Формула для АДИ

Универсальный рецепт для формулы в этот раз немного другой:

1. Выбрать статистику, которая будет слабо сходиться к какому-то распределению при $n \to \infty$
2. Найти выборочные квантили, соответствующие $\alpha = 0.05$ и другим параметрам выбранного распределения

3. Оценить статистику выборочными квантилями
4. Выразить искомую величину и получить результирующий АДИ

Про выбор статистики

Пожалуй, это самое сложное во всей лабе, учитывая, что в первом задании статистика дается в явном или в почти явном виде. Здесь же ее придется выводить самостоятельно на основании подсказок, приведенных в файле.
Рекомендую искать информацию по подсказкам в конспекте самого Ивана Александровича, т.к. там теоремы/свойства, которыми он предлагает воспользоваться, имеют такое же наименование, как и в задании к лабе (конспект можно найти в группе в треде "Материалы и объявления по теории").

После того, как получен АДИ

После вывода искомой формулы, остается только нагенерить выборок и протестировать АДИ на них. Если эмпирический процент попадания $\xrightarrow{n\to\infty} 95\%$, значит все сделано правильно. Если нет, значит ошибка скорее всего в формуле для АДИ.

Дополнительные полезные штуки

Ссылки

- Про доверительные интервалы
- Про разные тесты, большая часть которых и используется в первом задании
- Некоторые таблицы квантилей распределений. На самом деле, для получения необходимых квантилей можно воспользоваться SciPy или даже экселем — там есть для этого необходимые функции (и, возможно, это будет даже удобнее).

Для дочитавших

Надеюсь, сие руководство было полезным и существенно уменьшило ваше время, потраченное на эту лабу!
На составление гайда и самостоятельный разбор темы у меня уходит немало времени, поэтому я буду рад любой материальной поддержке! Если захочется меня поблагодарить, то можно кидать на сбер по номеру телефона (8-958-092-05-98).