ПСАД. ВШЭ. Задачи на семинар №2.

12-13 сентября

1 Простая задачка

Скачайте по ссылке данные. Считайте их в R командой `read.csv(<filename>, header = FALSE)`.

1. Посчитайте среднее каждого столбца
2. Посчитайте дисперсию каждого столбца
3. Посчитайте коэффициент корреляции Пирсона между двумя столбцами
4. Нарисуйте график.

2 Метод моментов

У Кати в сумке есть N ключей от квартиры. Когда Катя приходит домой после тяжёлого рабочего дня, она лезет в сумку, достаёт случайный ключ и пытается им открыть дверь. Если у неё не получается, она кладёт ключ обратно, случайно достаёт ещё один ключ и так далее пока дверь не откроется.

Пусть в течение пяти дней она смогла открыть дверь на 8-ую, 12-ю, 7-ую, 6-ю и 12-ую попытки соответственно. Найдите оценку максимального правдоподобия, а также оценку методом моментов на число N, число ключей в сумке у Кати. (Предполагается, что ключи одинаковые и Катя достаёт ключи равновероятно)

3 Метод максимального правдоподобия

Возобновляемые источники электроэнергии с каждым днем становятся все более и более востребованы человечеством не только в силу удорожания электроэнергии, производимой обычными средствами, но также и из-за щадящего режима использования природных ресурсов. Так, ветрогенератор мощностью 1 МВт сокращает ежегодные выбросы в атмосферу на 1800 тонн CO_2, 9 тонн SO_2 и 4 тонн оксидов азота. А по оценкам Global Wind Energy

Council к 2050 году мировая ветроэнергетика позволит сократить ежегодные выбросы CO_2 на 1.5 миллиарда тонн.

К недостаткам такого производства электроэнергии среди прочих можно отнести тяжелую прогнозируемость вырабатываемых мощностей. В частности, выработка энергии с помощью ветряков сильно зависит от переменчивой скорости ветра. Поэтому предсказание скорости ветра является очень нужной и важной задачей.

Для моделирования скорости ветра часто прибегают к распределению Вейбулла, которое зависит от двух параметров (масштаба λ, формы k) и имеет плотность

$$f(x \mid k, \lambda) = \frac{k x^{k-1}}{\lambda^k} \exp\left\{-(x/\lambda)^k\right\} \tag{1}$$

Если честно взять производные от функции правдоподобия по параметрам k, λ и положить $\frac{\partial L}{\partial k} = \frac{\partial L}{\partial \lambda} = 0$, то получим

$$\frac{1}{k} + \frac{\sum \log(x_i)}{n} - \frac{1}{\alpha} \sum x_i^k \log(x_i) = 0 \tag{2}$$

где $\lambda^k = \frac{1}{n} \sum_i x_i^k = \frac{\alpha}{n}$. К сожалению, найти точное решение относительно k не получится, поэтому мы будем пользоваться численными методами.

Вам необходимо

1. Скачать по ссылке данные со скоростью ветра для данной местности.
2. Численно решить задачу (2) с помощью функции `uniroot`;
3. построить график распределения Вейбулла с параметрами, которые вы нашли, и удостовериться, что эмпирические данные этому не противоречат. В этом вам поможет команда `hist` и команда `curve`;
4. ввести в гуглоформу в качестве ответа на это задание значение параметра λ.

4 Задача на бутстрэп

У ветрогенераторов нашей модели есть одна существенная проблема – они расчитаны на скорость ветра больше 5 м/с. При скорости ветра меньше 5 м/с они не могут эффективно вырабатывать электроэнергию, а если скорость преодолевает порог, работают примерно на фиксированной мощности. Поэтому доля времени когда скорость ветра преодолевает данный порог является ключевым критерием, определяющем рентабельность установки ветрогенератора на данной местности. Т.к. потенциальных кандидатов на место установки ветрогенератора может быть очень много, важно уметь эффективно и точно оценивать данную величину по нескольким замерам скорости ветра в разные моменты времени.

Существует два способа оценить вероятность того, что скорость ветра превышает 5 м/с:

- **Метод 1.** Посчитать долю измерений в выборке, значение которых превышает 5 м/с.
- **Метод 2.** Посчитать вероятность $p(x > 5)$ по распределению Вейбулла (см. Задание 2), параметры которого оценены при помощью метода максимального правдоподобия по выборке.

Для того, чтобы оценить эффективность каждого метода воспользуемся бутстрэпом.

Для этого N раз просемплируйте подвыборку $\boldsymbol{X}_i$ размера K исходной выборки с повторениями. По каждой из полученных подвыборок $\boldsymbol{X}_i$ посчитайте значение статистик T_1 и T_2 соответствующих каждому из двух методов. $t_i^{(1)} = T_1(\boldsymbol{X}_i)$, $t_i^{(2)} = T_2(\boldsymbol{X}_i)$. После чего посчитайте выборочное стандартное отклонение данной величины в завимисостиот подвыборки, по которой она была оценена. Чем меньше стандартное отклонение, тем эффективнее (точнее) метод.

В задании предлагается:

1. Оценить вероятность того, что скорость ветра превысит 5 м/с двумя способами. Сравнить результаты.
2. При помощи бутстрэпа оценить дисперсию обоих методов, а также построить 95 % доверительные интервалы для каждой оценки.

   Сравнить результаты.

5 Задача на дельта метод

В медицине, а также в некоторых других областях (например, связанных с азартными играми), помимо оценки вероятности успеха p какого-либо события, также часто используется понятие *шансов (odds)*, которое равно вероятности успеха деленной на вероятность неудачи, или

$$\text{odds} = \frac{p}{1-p} \tag{3}$$

Понятие шансов достаточно интуитивно и обычно его используют для придания вероятности некоторой наглядности, например, если $p_1 = 0.1$, то odds $= 1/9$, или, как часто говорят, «шансы 9 к одному против чего-либо» (проигрша, неуспеха и т.д.), что приводит в свою очередь к корректной частотной интерпретации вероятности – в среднем «успех» будет появляться только один раз за 9 «неудач».

Часто на практике возникает ситуация, когда нужно сравнить шансы между собой. В таком случае пользуются так называемым отношением шансов (odds ratio, OR)

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)} \tag{4}$$

Такая оценка представляет собой количественную меру наличия или отсутствия взаимовлияния признаков. То есть, если OR > 1 то, очевидно, $p_1(1-p_2) > p_2(1-p_1)$. Чтобы лучше понять практический смысл OR, ответьте на следующий

Вопрос. Обозначим группу курящих людей как A, а группу некурящих людей как B. Каждого человека из каждой группы наблюдают всю его жизнь. За достаточно длительный промежуток времени, обнаружилось, что доля людей заболевших раком легких в группе A (p_1) и в группе B (p_2) отличается. Предположите, какие могут быть отличия между p_1, p_2 и покажите, как это отличие сказыается на значении OR. Ваш ответ введите в гуглоформу второго семинара, доступную на страницу

Из теории известно, что если $\hat{\theta}$ есть оценка ММП и $\phi(\cdot)$ некоторая функция, задающая взимно однозначное отображение, то оценка ММП $\widehat{\phi(\theta)}$ в точности равна значению функции от оценки ММП $\hat{\theta}$, то есть $\widehat{\phi(\theta)} = \phi(\hat{\theta})$. В частности, поэтому для оценки OR обычно используется ММП.

Обычно в приложениях часто используют не саму величину OR, а величину $\log(\text{OR})$. В случае, когда используют такую оценку, особый интерес представляет изменение дисперсии оценки $\widehat{\log(\text{OR})}$ по сравнению с дисперсией оценки $\widehat{\text{OR}}$.

Данные, по которым можно сделать такую оценку, (данные можно скачать по этой ссылке), представляют собой результаты исследования, проведенного в 1989-1994 годах и направленного на проверку эффективности малой дозы аспирина для уменьшения доли инфарктов у относительно здорового населения. Прочитать эти данные можно с помощью команды `df <- read.csv("aspirin.csv", row.names = 1)`.

По этим данным вам предлагается

1. проверсти bootstrap оценку параметра OR сделав N $= 1000$ bootstrap выборок (обозначив в формуле (4) p_1 – долю тех людей, у которых случился удар и которые принимали аспирин, а за p_2 – долю людей, у которых случился удар и которые не принимали аспирин) ;

2. провести bootstrap оценку параметра $\log(\text{OR})$ сделав N $= 1000$ boostrap выборок;

3. вывести аналитическую формулу изменения дисперсии при применении функции $\log(\cdot)$;

4. проверить, что дельта метод работает, подставив в аналитическую формулу из предыдущего пункта оценку из пункта 1) и сравнив получившиеся число с дисперсией $\log(\text{OR})$;

5. численные результаты и ваши выводы введите в соответствующую секцию гуглоформы.