ML/DS

- описательная статистика (как распределены ваши данные, моды распределения, ожидание и дисперсия) - https://baguzin.ru/wp/opredelenie-srednego-znacheniya-varia/
- теория вероятности («предполагая биномиальное распределение, какова вероятность увидеть 5 платёжеспособных клиентов за 10 кликов?»)
- проверка гипотез (формирование базиса для A/B тестирования, T-тесты, дисперсионный анализ (ANOVA), критерий хи-квадрат)
- регрессии (линейность зависимости переменных, возможные источники отклонений, свойства наименьших квадратов)
- байесовский вывод (его преимущества и недостатки в сравнении с частотными методами).

Пример вопроса: Есть набор данных. Он содержит недостающие значения, которые распределены вдоль одного стандартного отклонения от медианы. Какой процент данных останется неизменным? Почему?

Ответ: В этом вопросе есть подсказка: так как данные распределены по медиане, можно предположить, что речь идет о нормальном распределении. Нам известно, что при нормальном распределении ~68% данных лежит в одном стандартном отклонении от медианы, а значит, ~32% данных остается неизменным. Таким образом, ~32% данных останется неизменным при недостающих значениях.

Метрики классификации

- Матрицы ошибок для измерения точности, полноты и чувствительности модели.
- F1 оценка.
- Истинно-положительный, истинно-отрицательный, ложно-позитивный и ложно-отрицательный результаты (TPR, TNR, FTR, FNR).
- Ошибки первого и второго рода.
- Кривые AUC-ROC.

Метрики регрессии

- Общая сумма квадратов, объяснённая сумма квадратов и сумма квадратов отклонений.
- Коэффициент детерминации и его скорректированная форма.
- Информационные критерии Акаике (AIC) и Байеса (BIC).
- Преимущества и недостатки погрешностей RMSE, MSE, MAE и MAPE.

Дилемма смещения-дисперсии, переобучение/недообучение

- Алгоритм k-ближайших соседей и подбор значения k в дилемме смещения-дисперсии.
- Случайные леса.
- Асимптотические свойства оценок.
- Проклятие размерности.

Эмпирическое правило

> Если данные имеют колоколообразное распределение, то приблизительно 68% наблюдений отстоят от математического ожидания не более чем на одно стандартное отклонение, приблизительно 95% наблюдений отстоят от математического ожидания не более чем на два стандартных отклонения и 99,7% наблюдений отстоят от математического ожидания не более чем на три стандартных отклонения.

Описательные статистики

Генеральная совокупность – это совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины.

Выборкой (выборочной совокупностью) называется совокупность случайно отобранных объектов из генеральной совокупности.

```
# нормальное распределение описывается с помощью среднего и отклонения
norm_rv1 = stats.norm(loc=35, scale = 10)
# scale - стандартное отклонение*
# loc - среднее*

#генерирует случайные значения из распредления norm_rv1*
gen_pop = norm_rv1.rvs(size=10000)
```

Медиана — это такое число выборки, что ровно половина из элементов выборки больше него, а другая половина меньше него. Робастная, то есть устойчивая, мера концентрации

Мода — значение во множестве наблюдений, которое встречается наиболее часто.

Квантиль — значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Если вероятность задана в процентах, то квантиль называется процентилем или перцентилем. 50ый перцентиль - медиана.

В нормальном распеделении среднее и медиана совпадают

Оценки генеральной совокупности

Для генеральной совокупности данных:

среднее -

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

дисперсия -

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

среднеквадратическое отклонение -

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Выборочные оценки

Зачастую у нас нет возможности работы с генеральной совокупностью, мы имеем дело только с выборкой, то есть мы не можем точно знать значения дисперсии и стандартного отклонения, поэтому данные показатели мы можем лишь оценить.

Среднее генеральной совокупности через **выборочное среднее**:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$$

, здесь n - объем выборки.

Стандартное отклонение через **выборочное стандартное отклонение**:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

.

Отличается от стандартного отклонения генеральной совокупности!

Разные выборки из генеральной совокупности могут отличаться от самой генеральной совокупности, то есть при оценке генеральной совокупности через выборку мы можем ошибаться. Такая ошибка называется **стандартная ошибка**. Рассчитывается для любого показателя, чаще всего используется **стандартная ошибка среднего** (т.е. оценивает точность, с которой выборочное среднее оценивает среднее генеральной совокупности).

Стандартная ошибка среднего в математической статистике — величина, характеризующая стандартное отклонение выборочного среднего, рассчитанное по выборке размера n из генеральной совокупности.

Чем больше выборка, тем точнее оценка среднего и тем меньше его стандартная ошибка. Чем больше изменчивость исходной совокупности, тем больше изменчивость выборочных средних, поэтому стандартная ошибка среднего возрастает с увеличением стандартного отклонения совокупности.

истинная стандартная ошибка -

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

оценка стандартной ошибки по выборке -

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

Центральная предельная теорема

- Выборочные средние имеют приближенно нормальное распределение независимо от распределения исходной совокупности, из которой были извлечены выборки.
- Среднее значение всех возможных выборочных средних равно среднему исходной совокупности.
- Стандартное отклонение всех возможных средних по выборкам данного объема, называемое стандартной ошибкой среднего, зависит как от стандартного отклонения совокупности, так и от объема выборки.

При этом:

- выборки должны извлекаться случайно
- размер выборки не должен превышать 10% размера всей генеральной совокупности
- размер выборки должен быть достаточно большим - принимают, что большая выборка - более 30 наблюдений

ЦПТ позволяет делать предположения о вашем исходном распределении.

Проверка гипотез

Пайплайн оценки статистической значимости:

- Формулирование нулевой гипотезы (H0);
- Оценка вероятности получить наблюдаемые (или более сильные) различия при условии справедливости нулевой гипотезы;
- Принятие либо отвержение нулевой гипотезы.

```
F, p = stats.f_oneway(sample_groups[0],sample_groups[1],sample_groups[2],sample_groups[3])
```

p - вероятность ошибочно отвергнуть верную нулевую гипотезу. Мы не сможем отвергнуть гипотезу H0, если p >0.05

Просто дисперсонный анализ проверяет что чем больше разброс средних и чем меньше разброс значений внутри групп, тем меньше вероятность того, что наши группы — это случайные выборки из одной совокупности.

Уровень значимости, ошибки первого и второго рода.

Уровень значимости — это такое (достаточно малое) значение вероятности события, при котором событие уже можно считать неслучайным.

В предыдущем примере мы установили уровень значимости (alpha) равным 0.05

Уровень значимости — допустимая для данной задачи вероятность ошибки первого рода (ложноположительного решения, false positive), то есть вероятность отклонить нулевую гипотезу, когда на самом деле она верна.

Ошибка первого рода — ситуация, когда отвергнута правильная нулевая гипотеза (англ. type I errors, α errors, false positive, ошибочное отвержение). **Ошибка второго рода** — ситуация, когда принята неправильная нулевая гипотеза (англ. type II errors, β errors, false negative, ошибочное принятие).

Критерий Стьюдента для несвязанных выборок

Частный случай дисперсионного анализа - применение критерия Стьюдента, позволяет проверять значимость различий двух групп.

t-критерий Стьюдента:

t =

$$\frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}} = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2_{\overline{X}_1} + S^2_{\overline{X}_2}}}$$

Для двух случайных выборок извлеченных из **одной** нормально распределенной совокупности это отношение будет близко к 0. Чем меньше t, тем больше вероятность нулевой гипотезы. Чем больше t, тем больше оснований отвергнуть нулевую гипотезу и считать, что различия статистически значимы.

```
t_stat = stats.ttest_ind(norm_sam_sample,norm_vel_sample)
t_stat
```

Нулевая гипотеза отвергается при 5-% уровне значимости, действительно, выборки взяты из разных распределений.

Условия применимости критерия:

```
- нормальность в распределении средних (?);
- равенство дисперсий выборок;
- независимость выборочных данных;
```

Критерий Стьюдента для связанных выборок

```
t = $\Large \frac{\overline{d}} {S_{\overline{d}}}$, где $\overline{d} $ - среднее изменений, $S_{\overline{d}}$ - стандартная ошибка.df_ =
```

```
df_ = pd.DataFrame()
df_['courier_id'] = np.array([1, 2, 3, 4, 5, 6])
df_['couriers_without_bike_time'] = np.array([39, 49, 47, 39, 28, 26])
df_['couriers_with_bike_time'] = np.array([29, 53, 43, 39, 33, 30])
df_.head()
```

H0: среднее время доставки курьеров на велосипеде и без велосипеда равны

H1: среднее время доставки курьеров на велосипеде не равны

```
#Корректно только в случае, если два распредления - одни и те же курьеры
stats.ttest_rel(df_.couriers_without_bike_time, df_.couriers_with_bike_time)
```

Статистическая мощность

Статистическая мощность - (реже "чувствительность") (англ. statistical power) - это вероятность того, что тот или иной статистический критерий правильно отклонит неверную нулевую гипотезу. Иными словами - это способность критерия обнаружить различия там, где они действительно существуют. (**FalseNegative**)

Качественные признаки

Качественные признаки - признаки не связаны между собой никакими арифметическими соотношениями, упорядочить их также нельзя. Единственный способ описания качественных признаков состоит в том, чтобы подсчитать число объектов, имеющих одно и то же значение. Кроме того, можно подсчитать, какая доля от общего числа объектов приходится на то или иное значение.

Cumulative Density Function (CDF) - «Какова вероятность того, что результат окажется меньше или равен такому-то?»

Probability Density Function (PDF) - вероятность функции распределения

Доверительные интервалы

t =

$$\frac{\text{Разность выборочных средних} - \text{Разность истинных средних}}{\text{Стандартная ошибка разности выборочных средних}}$$

t =

$$\frac{\overline{X_1} - \overline{X_2}}{S_{\overline{X_1} - \overline{X_2}}} = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_{\overline{X_1} - \overline{X_2}}}$$

ДИ для разности средних

$$(\overline{X_1} - \overline{X_2}) - t_{\alpha/2} S_{\overline{X_1} - \overline{X_2}} < \mu_1 - \mu_2 < (\overline{X_1} - \overline{X_2}) + t_{\alpha/2} S_{\overline{X_1} - \overline{X_2}}$$

Полученное неравенство задает **доверительный интервал** для разности средних, в этот интервал разность истинных средних попадет в 95% случаев, при a = 0.05.

```
#Расчет для задачи про курьеров на разных транспортных средствах
import statsmodels.stats.api as sms
cm = sms.CompareMeans(sms.DescrStatsW(norm_vel_sample), sms.DescrStatsW(norm_sam_sample))
print (cm.tconfint_diff())
```

То есть, с 95% вероятностью можно утверждать, что с новыми самокатами курьеры уменьшили время доставки от 2 до 11 минут.

ДИ для среднего

$$\overline{X} - t_{\alpha/2} S_{\overline{X}} < \mu < \overline{X} + t_{\alpha/2} S_{\overline{X}}$$

```
t = sms.DescrStatsW(norm_vel_sample)
t.tconfint_mean()
```

Out[4]:

`(33.20110922608528, 38.33652874013045)`

То есть истинное среднее с 95% вероятностью лежит в интервале от 33 до 38 минуты.

TFIDF

```
from sklearn.feature_extraction.text import CountVectorizer

# list of text documents
text = ["The quick brown fox jumped over the lazy dog."]

# create the transform
vectorizer = CountVectorizer()

# tokenize and build vocab
vectorizer.fit(text)

# summarize
print(vectorizer.vocabulary_)

# encode document
vector = vectorizer.transform(text)

# summarize encoded vector
print(vector.shape)
print(type(vector))
print(vector.toarray())
```

Литература

- Кобзарь. Прикладная математическая статистика (2006)
- Kanji. 100 statistical tests (2006)
- Глантц. Медико-биологическая статистика (1999)
- Лагутин. Наглядная математическая статистика (2007)