Основы статистики

Генеральная совокупность – множество всех объектов, относительно которых делаются выводы в рамках исследования.

Выборка – часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом).

Виды выборок:

- Простая случайная выборка (simple random sample)
- Стратифицированная выборка (stratified sample)
- Групповая выборка (claster sample)

Типы переменных

1. Количественные (numerical) – измеренные значения:
    - Непрерывные ([0; 1]);
    - Дискретные (1, 2,..).
2. Номинативные (categorical) – разделение на группы (1=м, 2=ж).
3. Ранговые (ordinal) – операции сравнения (распределение мест в забеге).

Виды графиков

Histogramm (гистограмма) – график, показывающий как часто значение переменной встречается на определенном промежутке.

Dot plot (точечный график) – график, в котором каждой точке соответствует одно значение выборки.

Box plot (ящик с усами) – график, показывающий медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. В ящик попадут значения (50% измерений), лежащих между квантилями $x_{0.25}$ и $x_{0.75}$. Вверх и вниз от ящика исходят два отрезка равные $1.5 \cdot (x_{0.75} - x_{0.25})$, то есть полтора межквартильных размаха. Точки, превышающие с своем отклонении полтора межквартильных размаха, отображаются отдельно.

Q-Q plot (график квантиль-квантиль) – показывает насколько выборочное значение соответствует нормальному распределению, линия – идеальное нормальное распределение.

Scatter plot (диаграмма рассеяния) – диагрмма, изображающая значения двух переменных в виде точек на декартовой плоскости.

Biplot – график первых двух компонент с вкладом каждой переменной.

Меры центральной тенденции

Мода (mode) – значение признака, котороые встречается максимально часто.

Медиана (median) – значение признака, которое делит упорядоченное множество данных пополам.

Среднее значение (mean) – сумма всех значений признака, деленная на количество измеренных значений.

Обозначения: M_x – среднее значение генеральной совокупности, $\overline{X}$ – среднее значение выборки.

Свойста среднего:

1. $M_x = \frac{1}{n} \sum x_i$
2. $M_{x+C} = M_x + C$
3. $M_{x \cdot C} = M_x \cdot C$
4. $\sum (x_i - M_x) = 0$

Меры изменчивости

Размах (range) – разность максимального и минимального значения.

Дисперсия (variance) – средний квадрат отклонений индивидуальных значений признака от их средней величины.

Среднеквадратическое отклонение (standard deviation, стандартное отклонение) – среднее отклонение индивидуальных значений признака от их средней величины.

Обозначения: D_x – дисперсия генеральной совокупности, σ – станадртное отклонение генеральной совокупности, sd_x – стандартное отклонение выборки.

Свойства дисперсии и стандартного отклонения:

1. $D_x = \frac{1}{n-1} \sum (x_i - \overline{x})^2$
2. $D_{x+C} = D_x$, $sd_{x+C} = sd_x$
3. $D_{x \cdot C} = D_x \cdot C^2$, $sd_{x \cdot C} = sd_x \cdot C$

Квантили распределения

Квантиль – значение, которое заданная случайная величина не превышает с фиксированной вероятностью: $P(X \le x_\alpha) \ge \alpha$.

Квартили – три значения признака, которые делят упорядоченное множество данных на четыре равные части.

Нормальное распределение

Нормальное распределение – унимодально, симметрично, отконения наблюдений от среднего подчиняются определенному вероятностному закону (правило 3σ):

1. $P(\bar{x} - \sigma < X < \bar{x} + \sigma) = 0.68$
2. $P(\bar{x} - 2\sigma < X < \bar{x} + 2\sigma) = 0.95$
3. $P(\bar{x} - 3\sigma < X < \bar{x} + 3\sigma) = 0.98$

Стандартизация (Z-преобразование) – преобразование полученных данных в стандартную Z-шкалу (Z-scores) со средним $M_Z = 0,\ D_Z = 1$:

$$z_i = \frac{x_i - \bar{X}}{sd_x}$$

Центральная предельная теорема

При многократном повторении эксперимента выборочные средние симметричным образов распределяться вокруг среднего значения генеральной совокупности, а стандартное отклонение такого распределения выборочных средних – стандартная ошибка среднего: $se_x = \frac{\sigma}{\sqrt{n}} = \frac{sd_x}{\sqrt{n}}$ при $n > 30$.

Доверительный интервал для среднего

$[\mu - 1.96\sigma,\ \mu + 1.96\sigma]$ – 95% всех выборочных средних включили бы в данный интервал среднее генеральной совокупности μ.
$[\mu - 2.58\sigma,\ \mu + 2.58\sigma]$ – 99% доверительный интервал.

Идея статистического вывода

Нулевая гипотеза H_0 – отсутствие значимых различий между средним значением выборки и средним значением генеральной совокупности.

Альтернативная гипотеза H_1 – значимое отклонение между средним значением выборки и средним значением генеральной совокупности.

p-уровень значимости – вероятность получения такого или еще более сильного отклонения от среднего значения, если верна H_0. Чем меньше p, тем больше оснований отклонить нулевую гипотезу. Обычно при $p < 0.05$ принимаем H_1, т.е. мы получили статистически значимое отклонение.

Ошибка 1 рода – приняли альтернативную гипотезу, хотя верна нулевая.

Ошибка 2 рода – приняли нулевую гипотезу, хотя верна альтернативная.

Распределение Стьюдента

Если число наблюдений невелико и σ неизвестно, то используется ***распределение Стьюдента*** (t-distribution): унимодально, симметрично, но наблюдения с большей вероятностью попадают за пределу $\pm 2\sigma$ от среднего значения M, чем при нормальном распределении.

Форма распределения определяется числом степеней свободы ($df = n - 1$, degrees of freedom). С увеличением df распределение стремится к нормальному.

Критерий Стьюдента

$$H_0 : M_1 = M_2 \qquad H_1 : M_1 \neq M_2$$

$$X_1 - X_2 \in t\,(df = n_1 + n_2 - 2) \qquad se = \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}},\ t = \frac{\overline{X_1} - \overline{X_2}}{se}$$

Зная число степеней свободы и t-значение, мы можем расчитать p-уровень значимости.

Применимость критерия Стьюдента:

1. Гомогенность дисперсий (приблизительно одинаковы), можно проверить используя критерий Левена или критерий Фишера.
2. Нормальность распределения при $n < 30$.

Проверка на нормальность

Тест Колмагорова-Смирнова и критерий Шапиро-Уилка: если получаем p-уровень значимости больше 0.05, значит наша выборка значимо не отличается от нормальной.

Критерий Манна-Уитни переходит к ранжиорванным значениям и может быть использован при наличии значительных выбросов в выборке.

Дисперсионный анализ

ANOVA, ANalysis Of VAriance – позволяет сранивать среднее значение трех и более групп.

$$H_0 : \ M_1 = M_2 = M_3 \qquad H_1 : \ !(M_1 = M_2 = M_3)$$

Мы говорим, что вся изменчивость наших данных (SST) может быть обусловлена изменчивостью внутри групп (SSW) и изменчивостью между группами (SSB).

Если $SSB \gg SSW$, то весьма вероятно что как минимум два средних значения отличаются друг от друга. Основной статистический показатель – критерий Фишера:

$$F = \frac{SSB}{m-1} \div \frac{SSW}{N-m},$$

где n – размер выборки, m – количество групп.

Попровка Бонферрони

Bonferroni correction – при увеличении количества групп, необходима корректировка значения p-уровня значимости. Необходимо уровень значимости разделить на количество парных сравнений в эксперементе: $\binom{m}{2} = \frac{m \cdot (m-1)}{2}$.

Критерий Тьюки

Tukey HSD – расчитываются доверительные интервалы разности между средними значениями групп. Является менее консервативным, чем поправка Бонферрони.

Многофакторный дисперсионный анализ

MANOVA, Multivariate analysis of variance – позволяет сранивать среднее значение трех и более групп в зависимости от нескольких переменных. Вся изменчивость обусловлена:

$$SST = SSW + SSB_A + SSB_B + SSB_A \cdot SSB_B$$

Корреляция

Коэффициент ковариации – мера линейной зависимости двух переменных:

$$cov_{XY} = \frac{\sum_i (x_i - \overline{X}) \cdot (y_i - \overline{Y})}{N-1}$$

Коэффициент корреляции Пирсона – показатель силы и направления взаимосвязи двух количественных переменных, знак показывает направление взаимосвязи:

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \cdot \sigma_Y} \in [-1;\ 1]$$

Коэффициент детерменации $R^2 = (r_{XY})^2 \in [0;\ 1]$ – показывает в какой степени дисперсия одной переменной обусловлена влиянием другой переменной.

Особенности корреляции:

- Коэффициент корреляции применим если взаимосвязь линейна и монотонна, а также при отсутствии значительных выбросов (иначе необходимо использовать непараметрические аналоги).
- Положительная или отрицательная корреляция не говорит о причинно-следственной зависимости между переменными.
- Корреляция между двумя переменными может обуславливаться существованием третьей переменной, влияющей на обе эти переменные.

Непараметрические аналоги коэффициент корреляции Пирсона

Коффициенты корреляции Спирмана и Кендалла, так же как и критерий Манна-Уитни, переходят от реальных значений переменных к ранжированным значениям.

Регрессионный анализ

Одномерный регрессионный анализ применяется для исследования взаимосвязи двух количественных переменных (независимая переменная – предиктор и зависимая переменная – критериальная). Изучает как одна переменная определяет, позваляется предсказать другую переменную.

Линия регрессии

Линия тренда задается уравнением $y = b_0 + b_1 x$, где b_0 – свободный член (intercept), который отвечает за значение y, где линия пересечет ось Y; b_1 – угловой коэффицент (slope).

Необходимо подобрать b_0 и b_1 так, чтобы линия максимально адекватно отображала связь данных переменных, при этом выдвигается гипотеза $H_0: \ b_0 = 0$.

Метод наименьших квадратов

Метод нахождения оптимальных параетров линейной регресии, таких, что сумма квадратов ошибок (остатков) была минимальна. Остаток – расстояние от реального значения до предсказаннного значения, лежащего на прямой.

$$b_1 = \frac{sd_Y}{sd_X} \cdot r_{XY} \qquad b_0 = \overline{Y} - b_1 \cdot \overline{X}$$

Коэффициент детерминации – доля дисперсии зависимой переменной Y, объясняемаая регресионной моделью:

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}},$$

где SS_{res} – (residuals) сумма квадратов остатков (расстояний до регрессионой прямой), а SS_{total} – общая изменчивость (сумма квадратов расстояний до прямой $y = \overline{Y}$). Таким образом, $R^2 \approx 1$ означает, что почти вся изменчивость переменной объясняется нашей регрессионной моделью.

Условия применимости:

1. Линейная взаимосвязь X и Y.

Если зависимость на самом деле нелинейна, то предсказание будет ошибочно. Пути ликвидации нелинейности:

- Трансформация Тьюки *(Tukey Ladder of Powers)* – возведение X в степень, теряется интерпретируемость.
- Логарифмическая трансформация (Log transformation) – взятие логарифма от X и/или Y, интерпретируемость коэффициента наклона b_1:

a) $\log Y = b_0 + b_1 \cdot \log X$ – на сколько процентов увеличится значение зависимой переменной при изменении зависимой переменной на один процент.
b) $\log Y = b_0 + b_1 \cdot X$ – при единичном изменении переменной X, переменная Y в среднем изменяется на $100 \cdot b_1$ процентов.
c) $Y = b_0 + b_1 \cdot \log X$ – изменение на 1% по X в среднем приводит к $0.01 \cdot b_1$ изменению по переменной Y.

- Трансформация Бокса-Кокса *(Box-Cox transformation)* – обычно используется для трансформации зависимой переменной в случае, если у нас есть ненормальное распределение ошибок и/или нелинейность взаимосвязи, а также в случае гетероскедастичности.

2. Независимость наблюдений.
    - Источники:
        a) Повторные измерения (на разных уровнях независимой переменной): снижение чувствительности теста, искуственное увелечение мощности теста (псевдореплекация).
        b) Повторные пробы (на одном и том же уровне независимой переменной): искажение результатов.
        c) Кластерезация данных (нет повторных измерений, но данные взяты из нескольких гомогенных групп): искажение результатов.
3. Независимость предикторов. Отсутствие мультиколлинеарности – линейной зависимости между предикорами.
    - Абсолютная мультиколлинеарность – корреляция между двумя предикторами равна ± 1.
    - Если мы хотим только предсказывать значения, то мультиколлинеарность не проблема.
    - Для выявления можно построить корреляционную матрицу.
    - VIF (*Variance Inflation Factor)* – показывает, насколько хорошо предиктор объясняется другими предикторами. Еси VIF > 10, то предиктор лучше исключить из модели. Квадртаный корень из VIF показывает, во сколько раз стала больше стандартная ошибка данного коэффициента, по сравнению с ситуацией, если он был независим от других предикторов.
4. Нормальное распределение остатков.
5. Гомоскедастичность – постоянная изменчивость остатков на всех уровнях независимой переменной.
    - Если мы построим регрессию, где зависимой переменной будет квадрат остатков модели $Y \sim X$, а независимой переменной будет предиктор X, и в этой модели окажется высокий и значимый R^2, это означает, что в данных

есть гетероскедастичность. Тест Бройша — Пагана (Breusch-Pagan test), тест Уайта (White test).

6. Отсутствие автокорреляции остатков.

Множественная линейная регрессия

Позволяет исследовать влияние сразу нескольких независимых переменных на одну зависимую переменную: $y = b_0 + b_1 x_1 + ... + b_n x_n$.

К условиям применимости добавляются проверка на мультиколлинеарность (сильная связь или идентичность некоторых независимых переменных) и нормальное распределение переменных (желательно).

Исправленный R^2 – скорректированный коэффициент детерменации. Рассчитывается при включении в модель дополнительных независимых переменных.

Смешнная регрессионная модель

Эффект – влияние независимой переменной, с помощью которой мы предсказываем зависимую переменную.

Фиксированный эффект *(main effect)* – влияние независимой переменной, представляющее основной интерес для исследователя.

Случайный эффект *(random mixed effect)* – влияние независимой переменной, не представляющее основной интерес для исследователя.

Задача классификации

Логистическая регрессия – исследование взаимосвязи между номинативной зависимой переменной, имеющей всего 2 градации, и различными независимыми переменными.

Кластерный анализ – решает задачу кластеризации, то есть для каждго наблюдения находит те наблюдения, которые очень похожи на него, и те, которые от него отличаются. При этом мы снижаем размерность данных.

Анализ номинативных данных

Проверка гипотезы о распределении номинативной переменной

H_0 – ожидаемое распределение, H_1 – распределение отлично от ожидаемого.

Наблюдаемые частоты O (*observed*), ожидаемые частоты E (*expected*).

Все наблюдения независимы.

Расстояние χ^2 Пирсона

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \in \chi^2(n-1)$$

Распределение χ^2 с k степенями свободы

Распределение суммы квадратов k независимых стандратных нормальных случайных величин ($\mu = 0,\ D = 1$).

Проверка гипотезы о взаимосвязи двух доминативных переменных

Когда две номинативные переменные типа “причина-следствие”.

H_0 - распределение не отличается от ожидаемого, H_1 – отлично от ожидаемого, иными словами существует взаимосвязь.

Ожидаемая таблица – распределение одного признака абсолютно одинаково на всех факторах другого признака.

Ячейки ожидаемой таблицы (ожидаемые частоты):

$f_{ij} = \frac{f_i \cdot f_j}{N}$, где f_i – сумма в строке, f_j – сумма в столбце, N – число измерений.

При расчете используем распределение $\chi^2((n-1)\cdot(m-1))$, где n – число столбцов, m – число строк.

Минимальное количество наблюдений в каждой ячейке должно быть больше 5.

Поправка Йетса

В теории χ^2 непрерывно, тогда как вычисляемые значения всегда дискретны, в результате H_0 может отвергаться слишком часто. Применяется, когда некоторые ожидаемые частоты меньше 10.

$$\chi^2_{Yates} = \sum^{k} \frac{(|f_o - f_e| - 0.5)^2}{f_e}$$

Точный критерий Фишера

Обычно используется, если нарушается одно из условий применимости критерия χ^2.

Логистическая регрессия

Главная зависимая переменная – номинативная с двумя градациями, а качестве предикторов могут быть так и номинативные, так и количественные переменные.

Переводим номинативную переменную в вероятность:

$$[0, 1] \ni p_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} \in [-\infty, +\infty]$$

Odds (шансы) - отношение вероятности успеха $(Y = 1)$ к вероятности неудачи $(Y = 0)$.

Odds всегда больше нуля, чтобы удовлетворить отрицательным занчениям в правой части, возьмем натуральный логарифм.

p	$1 - p$	$odds = \frac{p}{1-p}$	$\log odds = logit(p)$
0.2	$1 - 0.2 = 0.8$	$\frac{0.2}{0.8} = 0.25$	$\log 0.25 = -1.38$
0.5	$1 - 0.5 = 0.5$	$\frac{0.5}{0.5} = 1$	$\log 1 = 0$
0.8	$1 - 0.8 = 0.2$	$\frac{0.8}{0.2} = 4$	$\log 4 = 1.38$

Если $\log odds > 0$, значит $p > 1 - p$, а если $\log odds < 0$, то $p < 1 - p$, при этом $p = \exp(\beta_0 + \beta_1 x_1) / (1 + \exp(\beta_0 + \beta_1 x_1)$.

Intercept only model

H_0 – нормальное распределение описывает распределение коэффициентов логистической регрессии, $p = 1 - p$, $odds = 1$, $logit(p) = 0$, т.е. logit(p) имеет нормальное распределение со средним равным 0, тогда если разделить полученное среднее на стандартную ошибку, то получим z-value – расстояние до 0 в стандартных отклонениях. Зная свойства нормального распределения можем найти p-уровень значимости (вероятность) получить такое, или еще более сильное отклонение от 0 (двунаправленная гипотеза). Intercept в данном случае – логарифм шанса положительного исхода. Используя его мы можем найти $odds$, если $odds < 1$, значит вероятность успеха ниже, чем неудачи.

Модель с одним номинативным предиктором

Intercept – натуральный логарифм шансов положительного исхода для одной из градаций зависимой переменной. Коэффициент при номинативной переменной – логарифм отношения шансов положительного исхода одной градации независимой переменной к другой. В отличии от теста χ^2, логистическая регрессия не только указала что две переменные взаимосвязаны, но укаала шансы для разных градаций независимой переменной.

Модель с двумя номинативными предикторами

Включаем в модель сразу несколько номинативных переменных, включая их взаимодейтсвие. Intercept – натуральный логарифм шансов положительного исхода для одной из градаций независимой переменной для первой градации независимой

переменной и первой градации второй независимой переменной. Все остальные коэффициенты – сравнение шансов базового уровня с шансами после перехода.

Взаимодейтсвие двух предикторов – разность логарифмов отношения шансов рассчитанного для градаций одного из предикторов при разных градациях второго предиктора.

U -критерий Манна-Уитни

Статистический непараметрический критерий, используемый для оценки различий между двумя независимыми выбораками, в которых признак измерен в метрической или раноговой шкале.

Критерий Краскела-Уоллиса

Основная статистика – дисперия средних значений рангов в сраниваемых группах. При верности нулевой гипотезы распределение это статистики можно описать при помощи распределения χ^2 .

Кластерный анализ

Обучение без учителя, без обратной связи. Решает задачу кластеризации, то есть для каждго наблюдения находит те наблюдения, которые очень похожи на него, и те, которые от него отличаются. При этом мы снижаем размерность данных.

Метод k-средних

1. Сами решаем на сколько кластеров будем делить.
2. Случайно выбираем начальные позиции центроидов кластера.
3. Для каждого наблюдения определяем, к какому центроиду он ближе всего.
4. Обновим позиции центроидов (среднее по каждой переменной для группы).
5. Снова для каждого наблюдения определяем, к какому центроиду он ближе всего.
6. Если принадлежности некоторых точек изменились, то пункт 4, иначе алгоритм сошелся.

Визуализация. В методе существует элемент случайности. Если мы еще раз проведем кластерезацию, то возможно, некоторые точки попадут в другой кластер, т.к. наблюдения могут сгруппироваться иным образом.

Возможно метод сойдется не очень удачно: метод “увяз” в локальном минимуме. Решения: начальные точки брать наиболее далеко друг от друга; провести кластерный анализ много раз с разными начальными позициями.

Оптимальное число кластеров

Внутригрупповая сумма квадратов (within-cluster sum of squares) – сумма квадратов отклонений каждого наблюдения от центроида кластера.

Общая внутригрупповая сумма квадратов (total within-cluster sum of squares) – сумма внутригрупповых сумм квадратов каждого кластера.

Если добавление одного кластера в наши данные значительно понижает общую сумму квадратов, то в увелечении числа кластеров есть смысл.

Если при увелечении числа кластеров плавное снижение общей внтуригрупповой суммы квадратов, то значит нет явной класторной структуры в данных.

Метод иерархической кластерезации

Свободен от априорного предположения о количестве кластеров. Резуьтат – дерево кластеров (дендограмма). Постепенно объеденяет две самые близкие точки в кластер, заменея их центроидом.

Метод анализа главных компонент

Principal component analysis – в случае сильной корреляции регрессионная прямая может стать осью главной компоненты. Часть информации теряется, а размерность снижается.