Статистический анализ данных. Спецкурс.

Лекция 5. Проверка статистических гипотез

Ботанический сад-институт ДВО РАН

Кислов Д.Е.

21 ноября 2016 г.

Структура лекции

- Ошибки первого и второго рода. Общая структура статистических критериев;

Структура лекции

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;

Структура лекции

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;
- Проверка данных на соответствие нормальному распределению. Критерий Шапиро-Уилка;

Структура лекции

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;
- Проверка данных на соответствие нормальному распределению. Критерий Шапиро-Уилка;
- Непараметрические критерии. Критерий хи-квадрат. Критерий Колмогорова-Смирнова;

Структура лекции

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;
- Проверка данных на соответствие нормальному распределению. Критерий Шапиро-Уилка;
- Непараметрические критерии. Критерий хи-квадрат. Критерий Колмогорова-Смирнова;
- Дисперсионный анализ;

Структура лекции

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;
- Проверка данных на соответствие нормальному распределению. Критерий Шапиро-Уилка;
- Непараметрические критерии. Критерий хи-квадрат. Критерий Колмогорова-Смирнова;
- Дисперсионный анализ;
- Попарное сравнение в дисперсионном анализе (post-hoc анализ);

Проверка гипотез

Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой (H_0);

Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой (H_0);
- Ошибкой первого рода является вероятность отвергнуть H_0, когда она верна;

Проверка гипотез

Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой (H_0);
- Ошибкой первого рода является вероятность отвергнуть H_0, когда она верна;
- Ошибкой второго рода является вероятность принять H_0, когда верна альтернативная гипотеза;

Проверка гипотез

Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой (H_0);
- Ошибкой первого рода является вероятность отвергнуть H_0, когда она верна;
- Ошибкой второго рода является вероятность принять H_0, когда верна альтернативная гипотеза;

Обозначения

- Ошибка первого рода обозначается α;

Проверка гипотез

Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой (H_0);
- Ошибкой первого рода является вероятность отвергнуть H_0, когда она верна;
- Ошибкой второго рода является вероятность принять H_0, когда верна альтернативная гипотеза;

Обозначения

- Ошибка первого рода обозначается α;
- Ошбика второго рода обозначается β;

Проверка гипотез

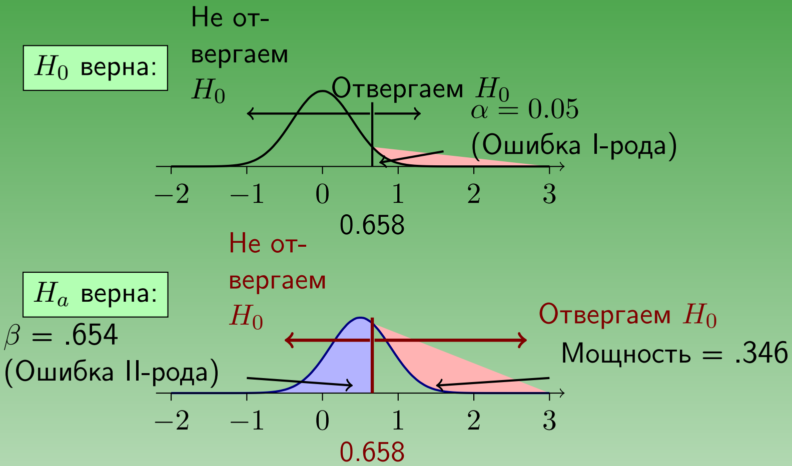
Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой (H_0);
- Ошибкой первого рода является вероятность отвергнуть H_0, когда она верна;
- Ошибкой второго рода является вероятность принять H_0, когда верна альтернативная гипотеза;

Обозначения

- Ошибка первого рода обозначается α;
- Ошбика второго рода обозначается β;
- $1 - \beta$ мощность критерия (вероятность принять альтернативную гипотезу, когда она верна);

Ошибки I-го и II-го рода. Мощность критериев.



Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Наиболее распространенные критерии согласия

- Критерий χ^2;

Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Наиболее распространенные критерии согласия

- Критерий χ^2;
- Критерий Колмогорова;

Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Наиболее распространенные критерии согласия

- Критерий χ^2;
- Критерий Колмогорова;
- Критерий Шапиро-Уилка;

Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Наиболее распространенные критерии согласия

- Критерий χ^2;
- Критерий Колмогорова;
- Критерий Шапиро-Уилка;
- другие специфичные критерии

Критерий χ^2 (К. Пирсон, 1900)

Формулировка критерия

Пусть $x_1, x_2, \ldots, x_N$ – выборочные данные. Проверяется гипотеза о том, что выборка получена из закона распределения $F(x)$. Тогда интервал возможных значений разбивается на k частей, в каждой из которых определяется число выборочных элементов. Далее, вычисляется статистика критерия:

$$\hat{\chi}^2 = N \sum_{i=1}^{k} \frac{(n_i/N - P_i)^2}{P_i},$$
$$P_i = F(b_{i+1}) - F(b_i)$$

b_i – границы разбиения. Величина $\hat{\chi}^2$ имеет распределение χ^2 с $k-1$ степенями свободы.

Если вычисленное $\hat{\chi}^2$ оказывается больше χ^2 квантиля для уровня α, то гипотеза о соответствии данному закону распределения отвергается при уровне значимости α.

Критерий Колмогорова

Формулировка критерия

Пусть проверяется соответствие выборочных данных $x_1, x_2, \ldots, x_N$ закону распределения $F(x)$. Исходя из выборочных данных формируется выборочная функция распределения $F_n(x)$

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leqslant x}$$

$$\forall t > 0\colon\ \lim_{n\to\infty} P(\sqrt{n}D_n \leqslant t) \approx K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

Критерий Колмогорова

Формулировка критерия

Пусть проверяется соответствие выборочных данных $x_1, x_2, \ldots, x_N$ закону распределения $F(x)$. Исходя из выборочных данных формируется выборочная функция распределения $F_n(x)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leqslant x}$$

$$\forall t > 0\colon\ \lim_{n\to\infty} P(\sqrt{n} D_n \leqslant t) \approx K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2}$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

Как быть если параметр распределения нужно оценивать по выборке?

Критерий Колмогорова

Формулировка критерия

Пусть проверяется соответствие выборочных данных $x_1, x_2, \dots, x_N$ закону распределения $F(x)$. Исходя из выборочных данных формируется выборочная функция распределения $F_n(x)$

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leqslant x}$$

$$\forall t > 0\colon\ \lim_{n\to\infty} P(\sqrt{n}D_n \leqslant t) \approx K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

Как быть если параметр распределения нужно оценивать по выборке?

Для случая нормального распределения используется поправка Лильефорса...

Критерий Колмогорова

Формулировка критерия

Пусть проверяется соответствие выборочных данных $x_1, x_2, \ldots, x_N$ закону распределения $F(x)$. Исходя из выборочных данных формируется выборочная функция распределения $F_n(x)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leqslant x}$$

$$\forall t > 0 \colon \lim_{n\to\infty} P(\sqrt{n} D_n \leqslant t) \approx K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

Для проверки однородности двух выборок применяют критерий Смирнова (построенный на базе распределения Колмогорова).

Критерий Шапиро-Уилка

Формулировка критерия

Статистика критерия имеет вид:

$$W = \frac{1}{s^2}\left[\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)\right]^2$$

$$s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2, \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Коэффициенты a_i находятся по таблицам; Критерий фактически основывается на отношениях оценок дисперсий распределения. Данный критерий является одним из самых мощных в плане проверки нормальности распределений.

Критерий Стьюдента. Предпосылки. Случай известной дисперсии.

Формулировка задачи

Пусть $x_1, x_2, \dots, x_n$ набор независимых нормально распределенных случайных величин с известной дисперсией σ^2. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу a. Нулевая гипотеза – среднее значение распределения есть a.

Решение

$$z = \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} (x_i - n \cdot a)$$

$H_1 = \{z < 0\}$ (или мат. ожидание меньше a): $z < \Phi_\alpha$ – нулевая гипотеза отвергается в пользу H_1 при уровне значимости α. Φ_α – α-квантиль стандартного нормального распределения.

Критерий Стьюдента. Предпосылки. Случай известной дисперсии.

Формулировка задачи

Пусть $x_1, x_2, \ldots, x_n$ набор независимых нормально распределенных случайных величин с известной дисперсией σ^2. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу a. Нулевая гипотеза – среднее значение распределения есть a.

Решение

$$z = \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} (x_i - n \cdot a)$$

$H_2 = \{z > 0\}$ (или мат. ожидание больше a): $z > \Phi_{1-\alpha}$ – нулевая гипотеза отвергается в пользу H_2 при уровне значимости α.

Φ_α – α-квантиль стандартного нормального распределения.

Критерий Стьюдента. Предпосылки. Случай известной дисперсии.

Формулировка задачи

Пусть $x_1, x_2, \ldots, x_n$ набор независимых нормально распределенных случайных величин с известной дисперсией σ^2. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу a. Нулевая гипотеза – среднее значение распределения есть a.

Решение

$$z = \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} (x_i - n \cdot a)$$

$H_3 = \{|z| > 0\}$ (или мат. ожидание не равно a): $z > \Phi_{1-\alpha/2}$ – нулевая гипотеза отвергается в пользу H_3 при уровне значимости α.

Φ_α – α-квантиль стандартного нормального распределения.

Критерий Стьюдента. Случай неизвестной дисперсии

Формулировка задачи

Пусть $x_1, x_2, \ldots, x_n$ набор независимых нормально распределенных случайных величин. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу a. Нулевая гипотеза – среднее распределения равно a.

Решение

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2,$$
$$m = \frac{1}{n}\sum_{i=1}^{n} x_i, t = \sqrt{n}\frac{m-a}{s}$$

$H_1 = \{t < 0\}$ (или мат. ожидание меньше a): $t < T_\alpha(n-1)$ – нулевая гипотеза отвергается в пользу H_1 при уровне значимости α.

Критерий Стьюдента. Случай неизвестной дисперсии

Формулировка задачи

Пусть $x_1, x_2, \ldots, x_n$ набор независимых нормально распределенных случайных величин. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу a. Нулевая гипотеза – среднее распределения равно a.

Решение

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2,$$
$$m = \frac{1}{n}\sum_{i=1}^{n} x_i, t = \sqrt{n}\frac{m-a}{s}$$

$H_2 = \{t > 0\}$ (или мат. ожидание больше a): $t > T_{1-\alpha}(n-1)$ – нулевая гипотеза отвергается в пользу H_2 при уровне значимости α.

Критерий Стьюдента. Случай неизвестной дисперсии

Формулировка задачи

Пусть $x_1, x_2, \ldots, x_n$ набор независимых нормально распределенных случайных величин. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу a. Нулевая гипотеза – среднее распределения равно a.

Решение

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2,$$
$$m = \frac{1}{n}\sum_{i=1}^{n} x_i, t = \sqrt{n}\frac{m-a}{s}$$

$H_3 = \{|t| > 0\}$ (или мат. ожидание не равно a):
$t > T_{1-\alpha/2}(n-1)$ – нулевая гипотеза отвергается в пользу H_3 при уровне значимости α.

Критерий Стьюдента. Сравнение двух средних.

Формулировка задачи

Пусть $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_m$ наборы независимых нормально распределенных случайных величин. Дисперсии выборок неизвестны, но равны. Необходимо проверить гипотезу о равенстве математических ожиданий, из которых получены эти выборки.

Решение

$$t = \frac{m_x - m_y}{s}\left(\frac{m \cdot n}{m+n}\right)^{1/2},$$

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m+n-2},$$

s_x^2, s_y^2 – несмещенные оценки дисперсий

$H_1 = \{t < 0\}$ $(m_x < m_y)$: $t < T_\alpha(n+m-2)$ – нулевая гипотеза отвергается в пользу H_1 при уровне значимости α.

Критерий Стьюдента. Сравнение двух средних.

Формулировка задачи

Пусть $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_m$ наборы независимых нормально распределенных случайных величин. Дисперсии выборок неизвестны, но равны. Необходимо проверить гипотезу о равенстве математических ожиданий, из которых получены эти выборки.

Решение

$$t = \frac{m_x - m_y}{s} \left(\frac{m \cdot n}{m + n} \right)^{1/2},$$

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m + n - 2},$$

s_x^2, s_y^2 – несмещенные оценки дисперсий

$H_2 = \{t > 0\}$ $(m_x > m_y)$: $t > T_{1-\alpha}(n + m - 2)$ – нулевая гипотеза отвергается в пользу H_2 при уровне значимости α.

Критерий Стьюдента. Сравнение двух средних.

Формулировка задачи

Пусть $x_1, x_2, \ldots, x_n,\ y_1, y_2, \ldots, y_m$ наборы независимых нормально распределенных случайных величин. Дисперсии выборок неизвестны, но равны. Необходимо проверить гипотезу о равенстве математических ожиданий, из которых получены эти выборки.

Решение

$$t = \frac{m_x - m_y}{s}\left(\frac{m \cdot n}{m+n}\right)^{1/2},$$

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m+n-2},$$

s_x^2, s_y^2 – несмещенные оценки дисперсий

$H_3 = \{|t| > 0\}$ $(m_x \neq m_y)$: $t > T_{1-\alpha/2}(n+m-2)$ – нулевая гипотеза отвергается в пользу H_3 при уровне значимости α.

Пример

Задача о сравнении

Даны два набора измерений длины плодов растений двух видов A и B. Значимы ли различия в средних значениях размеров плодов для этих видов.

Схема решения

- Проверка выборочных данных на соответствие нормальному распределению (Критерий Шапиро-Уилка, Колмогорова);
- Проверка равенства дисперсий (есть ли какие-либо основания, полагать дисперсии равными?);
- Выбор статистического теста (возможно критерий t-Стьюдента);

Задача

Задание на дом

Проведите оценку значимости различий средних значений для двух наборов линейных измерений, собранных для различных видов. Выбор видов и измеряемых параметров – исходя из Ваших научных интересов.

Непараметрический критерий сравнения

Гипотеза сдвига

Пусть $x_1, x_2, \dots, x_n$, $y_1, y_2, \dots, y_m$ наборы независимых случайных величин, полученных из законов распределения $f_1(x)$ и $f_2(x)$ соответственно.

$$H_0 = \{f_1(x) = f_2(x)\}$$
$$H_1 = \{f_1(x) = f_2(x - \mu)\}$$

При $n, m > 20$ выражение, основанное на U может быть приближено нормальным распределением.

Непараметрический критерий сравнения

Гипотеза сдвига

Пусть $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_m$ наборы независимых случайных величин, полученных из законов распределения $f_1(x)$ и $f_2(x)$ соответственно.

$$H_0 = \{f_1(x) = f_2(x)\}$$
$$H_1 = \{f_1(x) = f_2(x - \mu)\}$$

Критерий Манна-Уитни

$$U_x = mn + 0.5n(n+1) - \sum_i r(x_i)$$
$$U_y = mn + 0.5m(m+1) - \sum_j r(y_j) U = \min(U_x, U_y)$$

Если $U \in [U_1(\alpha), U_2(\alpha)]$, то H_0 отклоняется. При $n, m > 20$ выражение, основанное на U может быть приближено нормальным распределением.

Формулировка задачи

Имеется несколько наборов независимых случайных величин из нормального распределения с общей дисперсией σ. Требуется проверить гипотезу об одновременном равенстве всех средних этих выборок.

Задача

Сравниваются результаты морфометрических измерений какого-либо параметра для представителей вида в различных условиях произрастания. Нулевая гипотеза – различные условия не влияют на изменения этого параметра (т.е. средние для выборок равны). Альтернативная – влияние имеет место (средние для выборок различны).

Дисперсионный анализ

Критерий Фишера

$$x_{i,j} = M + (M_j - M) + (x_{i,j} - M_j), VAR_T = VAR_{WG} + VAR_{BG},$$
$$\frac{VAR_{BG}/(m-1)}{VAR_{WG}/(n-m)} > F_{m-1;n-m}(1-\alpha)$$

$F_{m-1;n-m}(1-\alpha)$– $1-\alpha$-квантиль распределения Фишера с $m-1$ и $n-m$ степенями свободы.

Схема провдеения дисперсионного анализа

Этапы параметрического ДА

- Проверка выборок на нормальность;
- Провера равенства дисперсий (критерий Левена, критерий Бартлетта (?));
- Вычисление статистики ДА, сравнение с критическим значением, или вычисление $p - value$;
- Если нулевая гипотеза отвергается, по необходимости проводится дополнительные исследования, с целью выяснить какие средние показали значимые различия. (тест Тьюки, тесты t-Стьюдента с поправками Бонферонни);