Прикладная статистика

и анализ данных

Съезд VII

Дисперсионный анализ I

Типы задач Д.А.

1. **Независимые выборки**

   Две группы пациентов. Одним дают одно лекарство, другим — другое. Верно ли, что первое лекарство эффективнее?

2. **Связные выборки**

   Пациент проходит испытание, принимает средство, затем снова проходит испытание. Отличается ли эффект?

- Методы для задач 2 типа можно использовать для задач 1 типа. При этом теряется важная информация.
- Методы для задач 1 типа *нельзя* использовать для задач 2 типа.

Независимые выборки

Человек	Препарат	Изменение температуры
Петя	Апотивадом	-0.9
Вася	Апотивадом	-0.6
Катя	Апотивадом	-1.0
Миша	Апотивадом	-0.3
Ира	Волымикер	-2.6
Света	Волымикер	-1.9
Коля	Волымикер	-0.7

Значимо ли отличается эффект от приема препаратов?

Связные выборки

Каждый человек применяет один и тот же препарат.

Человек	**Температура до**	**Температура после**
Петя	38.2	37.6
Вася	37.6	38.0
Катя	38.5	37.1
Миша	38.0	36.9
Ира	37.9	37.1
Света	39.4	37.3

Есть ли эффект от приема препарата?

Другие вопросы на практике

1. Значимо ли отличаются решения в топ-10 в соревновании на Kaggle?
2. На какой дизайн кнопки клиент кликнет с большей вероятностью?
3. Увеличился ли средний чек корзины покупателей после внедрения нового блока рекомендаций?
4. В чем причина оттока клиентов к конкурентам?
5. Отличаются ли гены по степени экспрессии?
6. многие другие...

Бернуллиевские выборки

Независимые выборки

$X_1, \ldots, X_n \sim Bern(p)$ и $Y_1, \ldots, Y_m \sim Bern(q)$.

$\mathrm{H}_0 : p = q$ *vs.* $\mathrm{H}_1 : p \; \{<, \neq, >\} \; q$

$\widehat{p}_1 = \overline{X} \overset{d}{\approx} \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$ и $\widehat{p}_2 = \overline{Y} \overset{d}{\approx} \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{m}\right)$ — ОМП

При справедливости H_0: $W(X, Y) = \frac{\widehat{p}_1 - \widehat{p}_2}{\widehat{\sigma}} \overset{d}{\approx} \mathcal{N}(0, 1)$,

где $\widehat{\sigma} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$.

Для $\mathrm{H}_1 : p > q$ критерий Вальда $S = \{W(x, y) > z_{1-\alpha}\}$.

Дов. интервал для $p_1 - p_2$ равен $C = (\widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha}\widehat{\sigma},\ 1)$.

H_0 отвергается $\Longleftrightarrow 0 \notin C$

Пример (влияние нового препарата на выздоровление)

Испытуемые делятся случайно на две группы:

1. *Исследуемая группа* — принимает новый препарат;

   $X = (X_1, \ldots, X_n) \sim Bern(p_1)$ — результаты лечения.

2. *Контрольная группа* — принимает плацебо;

   $Y = (Y_1, \ldots, Y_m) \sim Bern(p_2)$ — результаты лечения.

$\mathrm{H}_0: p_1 = p_2$ — отсутствие эффекта

$\mathrm{H}_1: p_1 > p_2$ — эффект присутствует

Пример (влияние нового препарата на выздоровление)

1. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 30$ человек, 21 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.7$

   $W(x, y) \approx 2$, $pvalue = 0.0228$, дов. интервал $(0.036, 1)$

2. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 30$ человек, 15 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.5$

   $W(x, y) \approx 3.76$, $pvalue = 0.00008$, дов. интервал $(0.225, 1)$

3. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 10$ человек, 7 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.7$

   $W(x, y) \approx 1.54$, $pvalue = 0.0618$, дов. интервал $(-0.017, 1)$

Связные выборки

$X_1, ..., X_n \sim Bern(p)$

$Y_1, ..., Y_n \sim Bern(q).$

$\mathrm{H}_0: p = q$ $vs.$ $\mathrm{H}_1: p \{<, \neq, >\}$ q

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	e	f
$X_i = 0$	g	h

Статистика **критерия**:

$$Z(X, Y) = \frac{\widehat{p} - \widehat{q}}{\sqrt{\frac{f+g}{n^2} + \frac{(f-g)^2}{n^3}}} = \frac{f - g}{\sqrt{f + g + \frac{(f-g)^2}{n}}} \xrightarrow{d_0} \mathcal{N}(0, 1)$$

Для $\mathrm{H}_1: p > q$ критерий $S = \{Z(x, y) > z_{1-\alpha}\}$

Дов. интервал для $p_1 - p_2$ равен $C = (\widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha}\widehat{\sigma},\ 1)$,

где $\widehat{\sigma} = \sqrt{\frac{f+g}{n^2} + \frac{(f-g)^2}{n^3}}$.

Пример

Два опроса с интервалом в полгода среди 1600 граждан Великобритании об одобрении премьер-министра.

	$Y_i = 1$	$Y_i = 0$	$\sum$
$X_i = 1$	$e = 794$	$f = 150$	944
$X_i = 0$	$g = 86$	$h = 720$	656
$\sum$	880	720	1600

H_0: рейтинг не изменился

H_1: рейтинг изменился $\Longrightarrow pvalue = 2.8 \cdot 10^{-5}$

H_1: рейтинг снизился $\Longrightarrow pvalue = 1.4 \cdot 10^{-5}$

H_1: рейтинг увеличился $\Longrightarrow pvalue = 0.99999$

Доверительный интервал $(0.0214, 0.0590)$

Дов. интервал Уилсона для $p - q$

Независимые выборки

$$[\hat{p} - \hat{q} - \delta,\ \hat{p} - \hat{q} + \varepsilon],$$

$$\delta = z_{\alpha/2}\sqrt{\frac{\ell_1(1-\ell_1)}{n} + \frac{u_2(1-u_2)}{m}},$$
$$\varepsilon = z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n} + \frac{\ell_2(1-\ell_2)}{m}},$$

ℓ_1, u_1 — корни уравнения

$$|x - \hat{p}| = z_{\alpha/2}\sqrt{\frac{x(1-x)}{n}},$$

ℓ_2, u_2 — корни уравнения

$$|x - \hat{q}| = z_{\alpha/2}\sqrt{\frac{x(1-x)}{m}}.$$

Связные выборки

$$[\hat{p} - \hat{q} - \delta,\ \hat{p} - \hat{q} + \varepsilon],$$

$$\delta = \sqrt{d\ell_1^2 - 2\hat{\varphi}\,d\ell_1 du_2 + du_2^2},$$
$$\varepsilon = \sqrt{du_1^2 - 2\hat{\varphi}\,du_1 d\ell_2 + d\ell_2^2},$$
$$\hat{\varphi} = \begin{cases} \frac{eh-fg}{(e+f)(g+h)(e+h)(f+g)}; \\ 0, \quad \text{если знаменатель} = 0; \end{cases}$$

$$d\ell_1 = \hat{p} - \ell_1, \quad d\ell_2 = \hat{q} - \ell_2,$$
$$du_1 = u_1 - \hat{p}, \quad du_2 = u_2 - \hat{q},$$

ℓ_1, u_1 — корни уравнения

$$|x - \hat{p}| = z_{\alpha/2}\sqrt{\frac{x(1-x)}{n}},$$

ℓ_2, u_2 — корни уравнения

$$|x - \hat{q}| = z_{\alpha/2}\sqrt{\frac{x(1-x)}{n}}.$$

Нормальные выборки

Связные выборки

$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$

$Y_1, ..., Y_n \sim \mathcal{N}(a_2, \sigma_2^2)$.

$\mathrm{H}_0 \colon a_1 = a_2$ *vs.* $\mathrm{H}_1 \colon a_1 \ \{<, \neq, >\} \ a_2$

Сведение к задаче с одной выборкой:

Рассмотрим выборку $\delta_1, ..., \delta_n$, где $\delta_i = X_i - Y_i$.

Тогда $\mathrm{H}_0 \colon \mathrm{E}\delta_i = 0$ *vs.* $\mathrm{H}_1 \colon \mathrm{E}\delta_i \ \{<, \neq, >\} \ 0$

Применяем критерий Вальда:

$$Z(X, Y) = \sqrt{n}\, \overline{\delta} / S_\delta \xrightarrow{d_0} \mathcal{N}(0, 1)$$

Почему не точный?

Если $X_i \sim \mathcal{N}(a_1, \sigma_1^2)$ и $Y_i \sim \mathcal{N}(a_2, \sigma_2^2)$ зависимы, то разность не обязана быть нормальной.

Связные выборки: обобщение

$X_1, \ldots, X_n$

$Y_1, \ldots, Y_n$.

$\mathrm{H}_0 \colon \mathrm{E}X_1 = \mathrm{E}Y_1$ *vs.* $\mathrm{H}_1 \colon \mathrm{E}X_1 \; \{<, \neq, >\} \; \mathrm{E}Y_1$

Сведение к задаче с одной выборкой:

Рассмотрим выборку $\delta_1, \ldots, \delta_n$, где $\delta_i = X_i - Y_i$.

Требование: $\delta_1, \ldots, \delta_n$ — выборка с конечной дисперсией.

Тогда $\mathrm{H}_0 \colon \mathrm{E}\delta_i = 0$ *vs.* $\mathrm{H}_1 \colon \delta_i \; \{<, \neq, >\} \; 0$

Применяем критерий Вальда:

$$Z(X, Y) = \sqrt{n}\, \overline{\delta} / S_\delta \xrightarrow{d_0} \mathcal{N}(0, 1)$$

Независимые выборки

$$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$$
$$Y_1, ..., Y_m \sim \mathcal{N}(a_2, \sigma_2^2).$$

Виды гипотез:

1. Равенство средних

   $H_0\colon a_1 = a_2 \quad vs. \quad H_1\colon a_1 \; \{<, \neq, >\} \; a_2$

   Способ зависит от доступной информации о дисперсиях.

2. Равенство дисперсий

   $H_0\colon \sigma_1 = \sigma_2 \quad vs. \quad H_1\colon \sigma_1 \; \{<, \neq, >\} \; \sigma_2$

3. Однородность

   $H_0\colon (a_1, \sigma_1^2) = (a_2, \sigma_2^2)$

Независимые выборки: среднее

$X_1, \ldots, X_n \sim \mathcal{N}(a_1, \sigma_1^2)$
$Y_1, \ldots, Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$.

$\mathrm{H}_0\colon a_1 = a_2$

$\mathrm{H}_1\colon a_1 \{<, \neq, >\} a_2$

Рассуждения:

$\overline{X} \sim \mathcal{N}\left(a_1, \sigma_1^2/n\right)$

$\overline{Y} \sim \mathcal{N}\left(a_2, \sigma_2^2/m\right)$

$\overline{X} - \overline{Y} \overset{\mathrm{H}_0}{\sim} \mathcal{N}\left(0, \sigma_1^2/n + \sigma_2^2/m\right)$

Случай 1. σ_1 и σ_2 известны

Статистика критерия

$$Z(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \overset{\mathrm{H}_0}{\sim} \mathcal{N}(0, 1)$$

Случай 2. $\sigma_1 = \sigma_2 = \sigma$ **неизвестны**

S_X^2, S_Y^2 — *несмещ.* оценки дисп.

Несмещенная оценка σ:

как взвешенное усреднение дисперсий

$$S_{tot}^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Статистика критерия

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{S_{tot}\sqrt{1/n + 1/m}} \overset{\mathrm{H}_0}{\sim} T_{n+m-2}$$

Случай 3. $\sigma_1 \neq \sigma_2$ и **неизвестны**

S_X^2, S_Y^2 — *несмещ.* оценки дисп.

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2/n + S_Y^2/m}} \overset{\substack{\mathrm{H}_0 \\ \text{прибл.}}}{\sim} T_v$$

$$v = \left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2 \Bigg/ \left(\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}\right)$$

Независимые выборки: дисперсия

$X_1, \ldots, X_n \sim \mathcal{N}(a_1, \sigma_1^2)$

$Y_1, \ldots, Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$.

$\mathrm{H}_0\colon \sigma_1 = \sigma_2$ *vs.* $\mathrm{H}_1\colon \sigma_1 \; \{<, \neq, >\} \; \sigma_2$

Критерий Фишера

S_X^2, S_Y^2 — *несмещенные* оценки дисперсии

$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{m-1}$

$$F(X, Y) = S_X^2 / S_Y^2 \overset{\mathrm{H}_0}{\sim} F_{n-1, m-1}$$

Не устойчив к отклонениям от нормальности даже асимптотически, нужна строгая проверка нормальности критерием Шапиро-Уилка.

Независимые выборки: однородность

$X_1, \ldots, X_n \sim \mathcal{N}(a_1, \sigma_1^2)$

$Y_1, \ldots, Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$

$\mathrm{H}_0 : (a_1, \sigma_1^2) = (a_2, \sigma_2^2)$

Идея:

1. Проверим гипотезу $\mathrm{H}_0' : \sigma_1^2 = \sigma_2^2$
   $\Longrightarrow$ *двусторонний* критерий Фишера;

2. Если не отвергается, то проверим $\mathrm{H}_0'' : a_1 = a_2$
   $\Longrightarrow$ *двусторонний* критерий Стьюдента (случай $\sigma_1 = \sigma_2$).

Как быть с уровнем значимости?

Применяем МПГ по методу Холма.


ВСЁ!