

Одномерный случай

Распределения

Нормальное распределение $N(a, \sigma^2)$ со средним a и дисперсией σ^2 . Свойства не перечисляю, надо знать. Также надо знать ЦПТ — центральную предельную теорему.

Распределение хи-квадрат $\chi^2(m)$ с m степенями свободы:

Пусть $\xi_i \sim N(0, 1)$, независимы. Тогда $\eta_m = \sum_{i=1}^m \xi_i^2$ имеет распределение $\chi^2(m)$. $\mathbb{E}\eta_m = m$, $\mathbb{D}\eta_m = 2m$.

Распределение $(\eta_m - m)/\sqrt{2m}$ сходится к $N(0, 1)$ при $m \rightarrow \infty$.

Распределение Стьюдента $t(m)$ с m степенями свободы:

Пусть $\xi \sim N(0, 1)$, $\eta_m \sim \chi^2(m)$, независимы. Тогда $\zeta_m = \xi/\sqrt{\eta_m/m}$ имеет распределение $t(m)$. $\mathbb{E}\zeta_m = 0$ ($m > 1$), $\mathbb{D}\zeta_m = m/(m-2)$ ($m > 2$).

Распределение ζ_m сходится к $N(0, 1)$ при $m \rightarrow \infty$.

Распределение Фишера-Снедекора $F(m_1, m_2)$ с m_1 и m_2 степенями свободы:

Пусть $\eta_1 \sim \chi^2(m_1)$, $\eta_2 \sim \chi^2(m_2)$, независимы. Тогда $\phi_{m_1, m_2} = \frac{\eta_1/m_1}{\eta_2/m_2}$ имеет распределение $F(m_1, m_2)$.

Соотношение между распределениями: $\xi^2 \sim \chi^2(1)$, $\zeta_m^2 \sim F(1, m)$, распределение $m_1\phi_{m_1, m_2}$ сходится к $\chi^2(m_1)$ при $m_2 \rightarrow \infty$.

Оценки

Пусть ξ — некоторая случайная величина с математическим ожиданием a и дисперсией σ^2 , $(x_1, \dots, x_n)$ — выборка объема n .

$\bar{x} = \sum_{i=1}^n x_i/n$ — несмещенная, состоятельная оценка a , $\mathbb{E}\bar{x} = a$, $\mathbb{D}\bar{x} = \sigma^2/n$. Если $\xi \sim N(a, \sigma^2)$, то $\bar{x} \sim N(a, \sigma^2/n)$. В произвольном случае ($\mathbb{D}\xi < \infty$), распределение $\sqrt{n}(\bar{x} - a)/\sigma$ стремится к $N(0, 1)$ (асимптотическая нормальность выборочного среднего).

$s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$ — асимптотически несмещенная, состоятельная ($\mathbb{D}\xi < \infty$) оценка σ^2 . $\tilde{s}^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$ называется исправленной выборочной дисперсией, $\sum_{i=1}^n (x_i - \bar{x})^2 = ns^2 = (n-1)\tilde{s}^2$. Если $\xi \sim N(a, \sigma^2)$, то $ns^2/\sigma^2 = (n-1)\tilde{s}^2/\sigma^2 \sim \chi^2(n-1)$.

Иногда пишут: $\tilde{s}^2 = SS/d.f.$, имея в виду, что SS — sum of squares (сумма квадратов), а d.f. — число степеней свободы (degree of freedom) соответствующего распределения χ^2 .

Расстояние в сигмах

В модели нормального распределения, расстояние от математического ожидания a измеряется в сигмах σ . ‘Далеко’ — то, что случается с маленькой вероятностью:

$\mathbb{P}(|\xi - a| > k\sigma) = 2(1 - \Phi(k)) = \alpha$, где Φ — функция распределения стандартного нормального распределения $N(0, 1)$.

Для $\alpha = 0.05$, $k = 1.96 \approx 2$, т.е. получаем правило двух сигм: вероятность попасть дальше, чем две сигмы, примерно равна 0.05 (мала). Аналогично, правило трех сигм и пр.

Гипотезы и критерии

Гипотеза $H_0 : \mathbb{E}\xi = a_0$

Статистика критерия (t-test): $t = \sqrt{n-1}(\bar{x} - a_0)/s = \sqrt{n}(\bar{x} - a_0)/\tilde{s}$. Если $\xi \sim N(a, \sigma^2)$, то t имеет распределение Стьюдента $t(n-1)$. Иначе асимптотическое распределение при $n \rightarrow \infty$ — нормальное распределение $N(0, 1)$. Также стандартное нормальное распределение получается в нормальной модели с известной дисперсией и $t = \sqrt{n}(\bar{x} - a_0)/\sigma$.

Заметим, что квадрат статистики критерия распределен как $\chi^2(n-1)$ и может быть записан в форме

$$t^2 = (\bar{x} - a_0) (\tilde{s}^2/n)^{-1} (\bar{x} - a_0) = r^2(\bar{x}, a_0).$$

Доверительный интервал в нормальной модели для $\mathbb{E}\xi$ имеет вид $\mathbb{P}(\mathbb{E}\xi \in (\bar{x} \pm c_\gamma \tilde{s}/\sqrt{n}) = \gamma$, где $t(n-1)(c_\gamma) = \gamma$. Также, доверительный интервал может быть записан в форме окрестности вокруг $\bar{x}$: $\mathbb{P}(r^2(\bar{x}, \mathbb{E}\xi) < c_\gamma^2) = \gamma$.

Гипотеза $H_0 : \mathbb{E}\xi_1 = \mathbb{E}\xi_2$, признаки зависимые (измерены на одних и тех же индивидах)

Данные имеют вид $(x_1, y_1)^T, \dots, (x_n, y_n)^T$. Эта гипотеза равносильна гипотезе $H_0 : \mathbb{E}(\xi_1 - \xi_2) = 0$, т.е. сводится к предыдущему варианту для выборки $z_i = x_i - y_i$.

Гипотеза $H_0 : \mathbb{E}\xi_1 = \mathbb{E}\xi_2$, признаки независимые (один признак, измеренный на разных индивидах)

Данные имеют вид $(x_1, \dots, x_{n_1})$ и $(y_1, \dots, y_{n_2})$. Соответственно, есть $\bar{x}$, $\bar{y}$, s_1^2 и s_2^2 , а также их исправленные варианты.

Для случая, когда предполагается, что $\mathbb{D}\xi_1 = \mathbb{D}\xi_2 = \sigma^2$, введем в качестве оценки σ^2 понятие pooled variance (объединённой дисперсии, pool — бассейн), когда из каждой выборки вычитается свое среднее, а потом они объединяются (сливаются). Введем ее исправленный вариант:

$$\tilde{s}^2 = ((n_1 - 1)\tilde{s}_1^2 + (n_2 - 1)\tilde{s}_2^2)/(n_1 + n_2 - 2) = \left(\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2 \right) / (n_1 + n_2 - 2)$$

Если $\xi_i \sim N(a_i, \sigma^2)$, то $(n_1 + n_2 - 2)\tilde{s}^2/\sigma^2 \sim t(n_1 + n_2 - 2)$.

Статистика критерия имеет вид

$$t = (\bar{x} - \bar{y}) / (\tilde{s} \sqrt{1/n_1 + 1/n_2}), \quad t^2 = (\bar{x} - \bar{y}) (\tilde{s}^2 (1/n_1 + 1/n_2))^{-1} (\bar{x} - \bar{y}). \quad (1)$$

Если верна модель $\xi_i \sim N(a_i, \sigma^2)$, то t , введенная в формуле (1) имеет распределение Стьюдента $t(n_1 + n_2 - 2)$, а $t^2 \sim F(1, n_1 + n_2 - 2)$. Иначе точное распределение неизвестно, но асимптотическое распределение при $n_1, n_2 \rightarrow \infty$ является стандартным нормальным.

Гипотеза $H_0 : \mathbb{E}\xi_1 = \dots = \mathbb{E}\xi_k$ для независимых выборок в нормальной модели

Если модель имеет вид $\xi_i \sim N(a_i, \sigma^2)$ (т.е. еще и дисперсии одинаковые), то проверка этой гипотезы называется дисперсионным анализом. Точнее — одномерным однофакторным дисперсионным анализом (1-way ANOVA, ANalysis Of VAriance). Данные обычно обозначаются как y_{ij} , где $i = 1, \dots, k$ — номер группы индивидов, на которых измеряется ξ_i , $j = 1, \dots, n_i$ — размеры групп, $n = \sum_{i=1}^k n_i$. Статистика критерия строится на основе разложения дисперсии (разложения суммы квадратов на выборочном языке) $SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$, где $SS_{\text{total}} = \sum_{ij} (y_{ij} - \bar{y})^2 \sim \chi^2(n-1)$, $SS_{\text{between}} = \sum_i n_i (\bar{y}_i - \bar{y})^2 \sim \chi^2(k-1)$, $SS_{\text{within}} = \sum_{ij} (y_{ij} - \bar{y}_i)^2 \sim \chi^2(n-k)$. Статистика критерия имеет вид $t = (SS_{\text{between}}/(k-1)) / (SS_{\text{within}}/(n-k)) \sim F(k-1, n-k)$.

Гипотеза $H_0 : \mathbb{D}\xi_1 = \dots = \mathbb{D}\xi_k$ для независимых выборок в нормальной модели

Стандартный критерий Фишера для $k = 2$ в нормальной модели: $t = \tilde{s}_1^2/\tilde{s}_2^2 \sim F(n_1 - 1, n_2 - 2)$. На гипотезу $H_0 : \mathbb{D}\xi_1 = \dots = \mathbb{D}\xi_k$ он обобщается как $t = \tilde{s}_{\min}^2/\tilde{s}_{\max}^2$ (и критическая область с одной стороны; распределение другое). Есть еще критерий Левена, который тоже естественным образом обобщается на случай k дисперсий.

Можно еще рассматривать критерий Бартлетта:

$t = A(n_1, n_2, k) \left((\sum_{i=1}^k n_i - k) \ln \tilde{s}^2 - \sum_{i=1}^k (n_i - 1) \ln \tilde{s}_i^2 \right)$ с асимптотическим распределением $\chi^2(k-1)$ при $\min n_i \rightarrow \infty$.

Задания

1. Выборочное среднее равно 0.8, исправленная выборочная дисперсия равна 4, объем выборки равен 25. Проверьте гипотезу, что мат.ожидание равно 2. Нужно найти p-level (вероятностный уровень) и сформулировать, при каких уровнях значимости гипотеза отвергается, при каких не отвергается.
2. Пусть $\eta \sim \chi^2(72)$. Можно ли приближенно считать расстояние от мат. ожидания в сигмах, почему? Если да, то примерно в скольких сигмах находится 108 от 72?
3. Проверить гипотезу, что скорость бега, в среднем, не зависит от настроения утром, если в 10 дней с хорошим настроением скорость при утренней пробежке была, в среднем, 12 км/ч со стандартным отклонением 0.2, тогда как за 20 дней с плохим настроением скорость бега была, в среднем, 13 км/ч со стандартным отклонением 0.5. Проверьте гипотезу, условно предположив, что модель нормальная и дисперсии одинаковые.
4. После года занятий в тренажерном зале у спортсмена возникла гипотеза, что за короткую тренировку давление не повышается. В результате измерения давления в течение 25 дней до и после тренировки было получено, что давление, в среднем, выросло на 1. При этом стандартное отклонение до тренировки было равно 4, а после тренировки - 6. Корреляция между измерениями давления оказалась равна 0.8. Проверьте возникшую гипотезу.
5. В трех группах с одинаковым числом студентов, равным 20, по оценкам на экзамене были сосчитаны $SS_{total} = 25$, $SS_{between} = 4$. Проверьте гипотезу, что уровень студентов во всех трех группах одинаковый.