Критерии согласия и проверка нормальности

Задача определения типа распределения возникает во многих приложениях (финансы, страхование и пр.). В страховании, например, важны распред типа Парето, в т. надёжности – Вейбулла и его част. случай нормальное распр. Критерии, основан. на нормальности данных, широко распростр. из-за 1) ЦПТ и 2) удобства вывода критериев.

Сначала рассм. задачу проверки гипотезы о том, что ~~распр.~~ истинное распределение – какое-то конкретное (напр., $N(0,1)$), а не класс распред. (напр., все нормал.)

Пусть есть выборка $X_1 \dots X_n$ – из неизв. распр. с ф.р. F.

Рассм. гипотезу H_0: $F = F_0$ и альтер. H_1: $F \neq F_0$. Критерии проверки таких гипотез назыв. критериями согласия.

1) Критерий Колмогорова – Смирнова.

Рассм. $D_n = \sup\limits_{x \in \mathbb{R}} |F_0(x) - \hat{F}_n(x)|$, где $\hat{F}_n(x) = \frac{1}{n}\sum\limits_{i=1}^{n} I(X_i \le x)$ – эмпирич. ф. р.

Т. Колмогорова Если F_0 – непр., то при верной гипотезе H_0

$$P(\sqrt{n} D_n \le x) \xrightarrow[n\to\infty]{} K(x) = \sum_{k \in \mathbb{Z}} (-1)^k e^{-2k^2x^2}$$

2) Критерий хи-квадрат Пирсона.

Разобьём $\mathbb{R}$ на 6–9 интервалов $\{B_i\}_{i=1}^{k}$ так, чтобы $P_0(B_i) \cdot n \ge 5$ $\forall i$ (n – размер выборки)

Обозн. $p_i^0 = P_0(B_i)$; $\mu_i = \#\{j : X_j \in B_i\}$, тогда

Т. Карла Пирсона $\hat{\chi} = \sum\limits_{i=1}^{k} \frac{(\mu_i - np_i^0)^2}{np_i^0} \xrightarrow{d} \chi^2_{k-1}$.

Сам критерий (и в крит. К–С аналогично): если $\hat{\chi} > u_{1-\alpha}$ (квантиль распр. χ^2_{k-1} уровня $1-\alpha$), то отвергаем H_0.

Здесь предпол. о непр F_0 не нужно.

Общее св-во, которым обладают критерии согласия – их статистика не зависит от F_0! Это очень важно для постр. критерия и помогает строить предельное распред. (иначе для каждого распред. оно было бы своим)

Однако ф-ции мощности на разных типах распред. будут разными. Кроме того, все критерии согласия – состоятельные.

Ещё критерии согласия:

- Смирнова – Крамера – фон Мизеса, со стат. $\int (F_n(x) - F(x))^2 dF(x)$
- Андерсона – Дарлинга (ω^2-крит.) $\int_{\mathbb{R}} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dF(x)$
- Фроцини $\int |F_n(x) - F(x)| \, dF(x)$
- Ватсона $\int \left[F_n(x) - F(x) - \int (F_n(x) - F(x)) dF(x) \right] dF(x)$

Иногда всё же исп. критерии, кот. зависят от вида распред.:

Джини: $\int_{\mathbb{R}} |F_n(x) - F(x)| \, dx$

Крамера – фон Мизеса $\int_{\mathbb{R}} (F_n(x) - F(x))^2 dx$

Критерии согласия и проверка нормальности

Критерии согласия хороши ещё и тем, что они работают для произвольных распределений, однако на практике применяют критерии, которые имеют бо́льшую мощность по сравнению с критериями согласия, но которые для других типов распред не работают (напр., критерий Шапиро-Уилка), но чтобы их использовать, надо сначала примерно понять, что за распред перед нами (с помощью графич. методов анализа).

- Есть общий метод проверки гипотезы $H_0: P \in \mathcal{P}_0 = \{P_\theta, \theta \in \Theta\}$ vs $H_1: P \notin \mathcal{P}_0$:

Берём некие состоят. оценки методом макс. пр-я (т. Фишера) параметров $\theta = (\theta_1 \dots \theta_d)$, тогда статистика критерия хи-квадрат: $\chi^2 = \sum_{i=1}^{k} \frac{(\mu_i - np_i)^2}{np_i} \to \chi^2_{k-1-d}$

при нек. усл. на дифф. $p_i(\theta_j)$

лучше так: $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum \frac{(\mu_i - np_i(\theta))^2}{np_i(\theta)}$

считается с наш распред. $P_{\hat{\theta}}$, где $P_{\hat{\theta}} \in \mathcal{P}_0$ и $\hat{\theta}$ – наша оценка

Проверка нормальности

- Так же мерой поступать и с крит. Колмогорова-Смирнова

Проверяем гипотезу $H_0: P \in \{N(a, \sigma^2)\}$ против $H_1: P \notin \{N(a, \sigma^2)\}$

Нужно взять статистику $D_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - \Phi(x)|$, где

F_n^* – эмпир. ф.р., $\Phi(x)$ – ф.р. $N(\bar{x}, s^2)$, где $\bar{x}$ и s^2 – выборочное среднее и дисперсия

Тогда $\sqrt{n} D_n$ сходятся к так наз. распред. Колмогорова.

Недостатки: имеет низкую мощность; если хвосты отличны от нормальных, то не различает это.

- Критерий Шапиро-Уилка (самый лучший)

$H_0: X \sim N(\mu, \sigma^2)$; $H_1: X \not\sim N(\mu, \sigma^2)$

Статистика $W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$, где $(a_1 \dots a_n) = \frac{m^T V^{-1}}{\|m^T V^{-1}\|^2}$, где

$m = (m_1 \dots m_n)^T$ – матожид. порядковых статистик выборки из $N(0,1)$,

V – их ковар. матрица.

W при верной H_0 имеет таблич. распред., значения $(a_1 \dots a_n)$ также табулированы.

Есть ещё критерий К² Д'Агостино, он тоже хороший, но похуже, чем Шапиро-Уилка. Он основан на статистиках skewness и kurtosis (выборочная асимметрия и выборочный эксцесс (почти)).

Есть более простой критерий, основ. на этих статистиках:

Критерии согласия и проверка нормальности

Критерий Жарка-Бера (в рус.-яз. литре ещё изв. как Харке-Бера)

$JB = \frac{n}{6}\left((Sk)^2 + \frac{1}{4}K^2\right)$, где $Sk = \frac{m_3}{(m_2)^{3/2}}$; $K = \frac{m_4}{(m_2)^2} - 3$, где

$m_j = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^j$.

При $n > 2000$ при верной H_0: $JB \xrightarrow{d} \chi^2(2)$. При малых n следует искать квантиль $u_{1-\alpha}$ с пом. моделирования.

Сам крит., естеств., такой: отвергать H_0, если $JB > u_{1-\alpha}$.

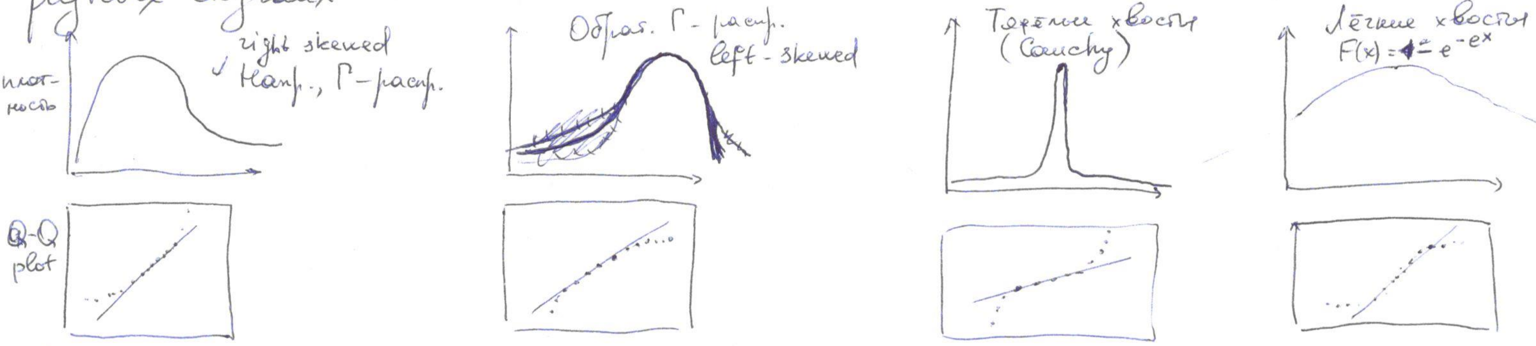
Есть ещё QQ-plots и PP-plots (последний реже используется)

Проверяем гипотезу H_0: $F = F_0\left(\frac{x-\mu}{\sigma}\right)$ (т.е. $\exists \in$ семейству распред с пар-рами сдвига и масштаба, например, $N(\mu, \sigma^2)$)

Делается след.: на график наносятся точки $\left(X_{(i)}, F_0^{-1}\left(\frac{i-0{,}5}{n}\right)\right)$, $i = 1 \dots n$. (Т.к. $F_0\left(\frac{X_1-\mu}{\sigma}\right)$ будет иметь распред $U[0,1]$ (известн. задача), порядк. статистики нашей выборки то $F_0(X_1)$ будет иметь распр. $U(\mu, \sigma+\mu)$ т.е. квантили этого распред. и коэф. у точек $\to \frac{i}{n}$ будут лежать на прямой. А $EX_{(i)} = u_{\frac{i}{n}}$, где $u_{\frac{i}{n}}$ – квантиль $U(\mu, \sigma+\mu)$)

Почему именно так, см. стр. 4

На примере норм. закона посмотрим, как ведут себя QQ-plots в разных случаях



Общие выводы о проверке нормальности:

1) Если выборка мала, то ничего понять нельзя, и по графикам тоже.
2) Если выборка большая, то $\forall$ крит. может выявить небольшие статистич. различия, кот. не явл. значимыми для практики.
3) Критерий Лиллиефорса представ. только историч. ценность. Лучше Шапиро-Уилка
4) Критерий хи-квадрат тоже не самый мощный и довольно общий, но статистиками используется.

Вопрос: а что делать с дискрет. распределением?

1) Крит. хи-квадрат; 2) сделать свёртку с каким-нибудь непр. ядром (см. оценку плотности) – и тогда будет непр. распред.!

Для проверки гипотезы $H_0: p = p_0$ в распр. Бернулли можно брать статистику

$$\frac{\bar{X} - p_0}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}} \xrightarrow{d} N(0,1)$$ по ЦПТ.

Критерии согласия и проверка нормальности

Теорема Фишера о χ^2-критерии

Итак, проверяем сложную гипотезу $H_0: P \in \mathcal{P} = \{P_\theta, \theta \in \Theta\}$, $\dim \Theta = d$ против альтернативы $H_1: P \notin \mathcal{P}$.

Как и ранее, разбиваем числ. ось на промежутки $\{B_i\}_{i=1}^k$, $\bigsqcup B_i = \mathbb{R}$.

$\mu_i = \#\{X_j : X_j \in B_i\}$; $p_i(\theta) = P_\theta(B_i)$, где $k > d$

Теорема Фишера

Пусть Θ - откр. мн-во в $\mathbb{R}^d$. Пусть выпол. условия

1) $\forall \theta \in \Theta$ ~~$\sum_{i=1}^k p_i(\theta) = 1$~~, $p_i(\theta) > c > 0$ $\forall i$

2) $\forall \theta \in \Theta$ $p_i(\theta)$ дважды непр. дифф. $\forall i$

3) $\forall \theta \in \Theta$ матрица $\left\| \frac{\partial p_i(\theta)}{\partial \theta_j} \right\|_{\substack{i=1\dots k \\ j=1\dots d}}$ имеет ранг d.

Пусть $\hat\theta$ - оценка, найд. методом макс. пр-я по выборке $\mu_1 \dots \mu_k$, т.е. $\hat\theta = \arg\max\limits_{\theta \in \Theta} L(\{\mu_i\}, \theta)$, где $L(\{\mu_i\}, \theta) = \frac{n!}{\mu_1! \dots \mu_k!} \prod\limits_{i=1}^k p_i^{\mu_i}(\theta)$.

или $\hat\theta$ - оценка по минимуму хи-квадрат:

$$\hat\theta = \arg\min_{\theta \in \Theta} \sum_{i=1}^k \frac{(\mu_i - n p_i(\theta))^2}{n p_i(\theta)}.$$ Тогда если H_0 верна, то

$$\chi^2(\hat\theta) = \sum_{i=1}^k \frac{(\mu_i - n p_i(\hat\theta))^2}{n p_i(\hat\theta)} \xrightarrow[n\to\infty]{d} \chi^2_{k-d-1}$$

Объяснение, почему работает Q-Q plot.

Вроде как точки $\left(X_{(i)}, F_0^{-1}\left(\frac{i-0{,}5}{n}\right)\right)$ должны лежать $\approx$ на одной прямой, если X_i - выборка из $F_0\left(\frac{x-\mu}{\sigma}\right)$.

Итак, пусть $Y_1 \dots Y_n$ - выборка из распред. с ф.р. F_0, тогда

$Y_{(i)} \xrightarrow[p\to 0]{p} F_0^{-1}\left(\frac{i}{n}\right)$, по т. о выборочн. квантилях, т.е. можно считать, что $Y_{(i)} \approx F_0^{-1}\left(\frac{i-0{,}5}{n}\right)$

Но $\frac{X_i - \mu}{\sigma} \stackrel{d}{=} Y_i$! Отсюда следует, что $\frac{X_{(i)} - \mu}{\sigma} \stackrel{d}{=} Y_{(i)}$,

т.е. точки $\left(X_{(i)}, F_0^{-1}\left(\frac{i-0{,}5}{n}\right)\right)$ должны лежать недалеко от прямой $\frac{x-\mu}{\sigma} = y$.