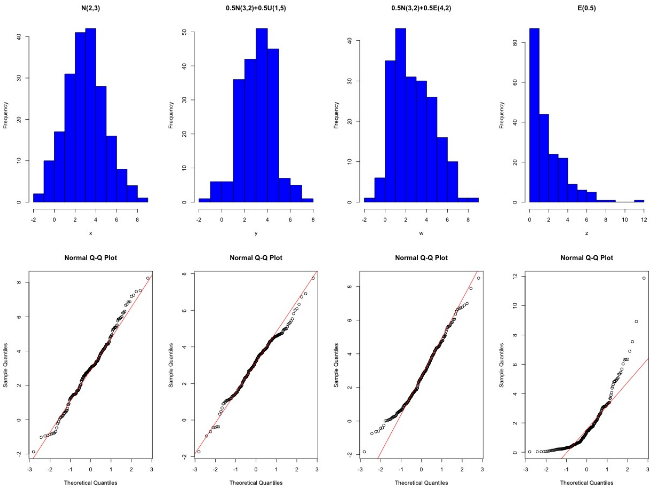
Критерии нормальности. Критерии однородности (продолжение)

Грауэр Л.В.

QQ-графики


N(2,3)
0.5N(3,2)+0.5U(1,5)
0.5N(3,2)+0.5E(4,2)
E(0.5)
Frequency
x
y
w
z
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

Критерии нормальности

ξ, $X_{[n]} = (X_1, \ldots, X_n)$

H_0: $\xi \sim N(a, \sigma)$

H_1: ξ имеет иное распределение.

Критерий Жарка-Бера

Критерий Лиллиефорса

Критерий Андерсона-Дарлинга

Критерий Шапиро-Уилка

Критерий Жарка-Бера

$$JB = \frac{n}{6}\left(Sk^2 + \frac{1}{4}K^2\right)$$

$$Sk = \frac{\hat{\mu}_3}{\tilde{s}^3}, \quad \tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

$$K = \frac{\hat{\mu}_4}{\tilde{s}^4} - 3, \quad \hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Если верна H_0, то

$$JB \xrightarrow[n\to\infty]{d} \zeta \sim \chi^2(2)$$

При больших объемах выборки

$$V_k = \{JB > \chi^2_{1-\alpha}(2)\}$$

При малых объемах выборки

$V_k = \{JB > C_{1-\alpha}\}$, $C_{1-\alpha}$ находят методом Монте-Карло.

Критерий Лиллиефорса

$$D(X_{[n]}) = \sup_{x \in \mathbb{R}} \left| F_n^*(x, X_{[n]}) - \Phi(x) \right|$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi} s} \int_{-\infty}^{x} e^{-\frac{(y-\bar{x})^2}{2s^2}} dy$$

$V_k = (C_{1-\alpha}, +\infty)$, $C_{1-\alpha}$ находят методом Монте-Карло

Критерий Андерсона-Дарлинга

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln F(X_{(i)}) + \ln(1 - F(X_{(n+1-i)})) \right]$$

$$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi} s} \int_{-\infty}^{x} e^{-\frac{(y-\bar{x})^2}{2s^2}} dy$$

$V_k = (C_{1-\alpha}, +\infty)$, $C_{1-\alpha}$ находят с помощью таблиц

Критерий Шапиро-Уилка

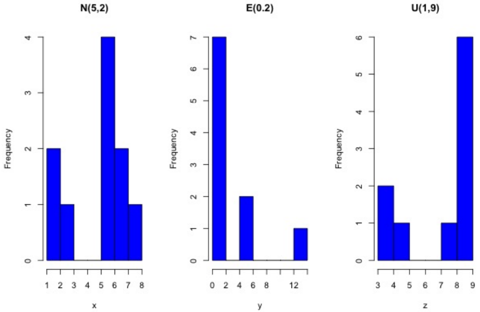
$$W = \frac{\left[\sum_{i=1}^{n} a_i X_{(i)}\right]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$n \leq 5000$

$W \in [0, 1]$

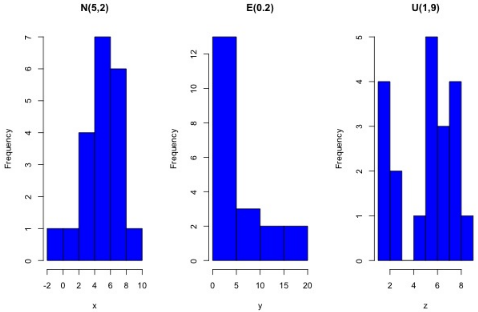
$V_k = [0, W_\alpha)$

N=10



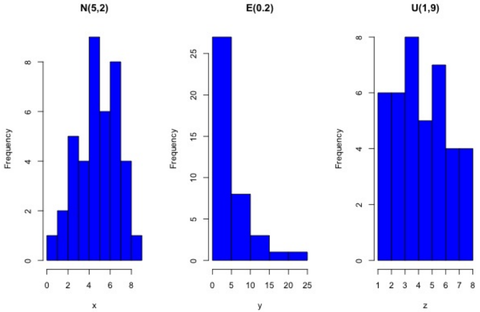
	JB	D	A^2	W
α	0.008	0.048	0.049	0.044
γ_1	0.157	0.293	0.427	0.455
γ_2	0.003	0.063	0.08	0.089

N=20



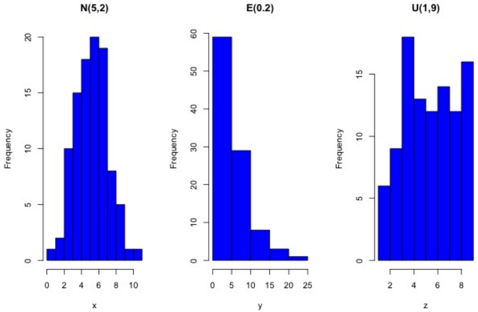
	JB	D	A^2	W
α	0.025	0.05	0.049	0.048
γ_1	0.472	0.567	0.782	0.836
γ_2	0.001	0.098	0.172	0.203

N=40



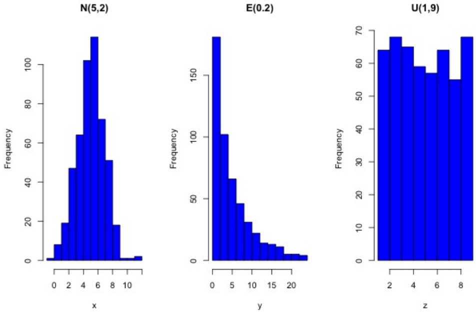
	JB	D	A^2	W
α	0.03	0.051	0.053	0.052
γ_1	0.87	0.905	0.98	0.99
γ_2	0.001	0.198	0.44	0.58

N=100



	JB	D	A^2	W
α	0.043	0.049	0.044	0.045
γ_1	1.0	1.0	1.0	1.0
γ_2	0.58	0.60	0.95	0.99

N=500



	JB	D	A^2	W
α	0.054	0.055	0.054	0.057
γ_1	1.0	1.0	1.0	1.0
γ_2	1.0	1.0	1.0	1.0

Критерий однородности Колмогорова–Смирнова

$\xi,\ F(x),\ X_{[n]} = (X_1, \dots, X_n)$

$\eta,\ G(x),\ Y_{[m]} = (Y_1, \dots, Y_m)$

$F(x)$ и $G(x)$ непрерывны

$$H_0:\ F(x) = G(x) \text{ для всех } x \in \mathbb{R}$$

$$H_1:\ F(x) \neq G(x) \text{ для всех } x \in \mathbb{R}$$

Теорема

$$D_{m,n} = \sup_{x \in \mathbb{R}} \left| G_m^*(x, Y_{[m]}) - F_n^*(x, X_{[n]}) \right| .$$

Если $F_0(x) = F(x) = G(x)$ непрерывна, тогда

$$P_0 \left\{ \sqrt{\frac{mn}{m+n}} D_{m,n} < z \right\} \longrightarrow K(z) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 z^2} .$$

$$V_k = \{ \sqrt{\tfrac{mn}{m+n}} D_{m,n} > d_{1-\alpha} \}$$

$$D_{m,n} = \sup_{x \in \mathbb{R}} \left| G_m^*(x, Y_{[m]}) - F_n^*(x, X_{[n]}) \right|$$

Критерий однородности хи-квадрат

$\xi_1, \xi_2, \ldots, \xi_k$

$\xi_j,\ F_j(x),\ X^j_{[n_j]} = (X^j_1, \ldots, X^j_{n_j}),\ j = 1, \ldots, k$

$$H_0 : \ F_{\xi_1} = \ldots = F_{\xi_k} = F_\xi \text{ для всех } x \in \mathbb{R}$$

$$H_1 : \ \exists F_{\xi_i} \neq F_{\xi_j}$$

Разобьем множество значений ξ_1, $\xi_2, \ldots, \xi_k$ на

$\Delta_1, \Delta_2, \ldots \Delta_r$: $\quad \Delta_i = (a_{i-1}, a_i]$, $i = 1, ..., r$

$$p_i = P\{\xi_1 \in \Delta_i | H_0\} = \ldots = P\{\xi_k \in \Delta_i | H_0\}$$

$$n_{ij} = num\{X_s^j \in \Delta_i\}$$

Статистика критерия

$$Z_j = \sum_{i=1}^{r} \frac{n_j}{p_i} \left(\frac{n_{ij}}{n_j} - p_i \right)^2 = \sum_{i=1}^{r} \frac{(n_{ij} - n_j p_i)^2}{n_j p_i}$$

$$Z = \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{(n_{ij} - n_j p_i)^2}{n_j p_i}$$

$$\hat{p}_i = \frac{\nu_i}{n}, \quad \nu_i = \sum_{j=1}^{k} n_{ij}, \quad i = 1, \ldots, r$$

$$\chi^2 = n \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{(n_{ij} - n_j \nu_i / n)^2}{n_j \nu_i}$$

Критическая область и p-value

Если H_0 верна, тогда

$$\chi^2(X_{[n]}) \xrightarrow[n\to\infty]{\mathrm{d}} \zeta \sim \chi^2((r-1)(k-1))$$

$\alpha \in (0,1)$

$$V_k = \{\chi^2(X_{[n]}) > \chi^2_{1-\alpha}((r-1)(k-1))\}$$

$$p - value = 1 - F_{\chi^2}(\chi^2(X_{[n]}))$$

Пример

Цена	Неважна	Мало важна	Важна	Очень важна
М	3	6	85	56
Ж	6	9	60	75

$n = 300$

Цена	Неважна или Мало важна	Важна	Очень важна
М	9	85	56
Ж	15	60	75

$\hat{p}_1 = 0.08,\ \hat{p}_2 = 0.48,\ \hat{p}_3 = 0.44$

$\chi^2(X_{[n]}) = 8.57$

$\chi^2_{0.95}((2-1)(3-1)) = 5.99$

* from statsci.org