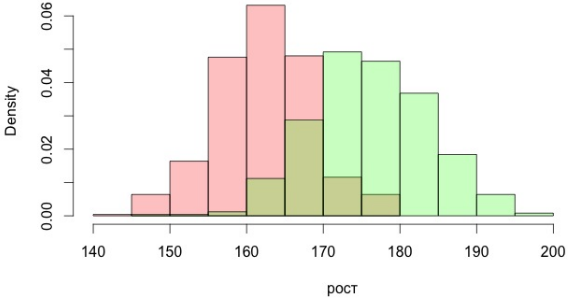
Стратификация

Грауэр Л.В.

Стратификация

Рост населения
Женщины 54%, Мужчины 46%



Примеры

- совокупность людей можно сгруппировать в страты по географической принадлежности
- пользователей интернета — по используемому браузеру
- финансовые транзакции — по величине транзации: большая, маленькая, средняя

Смесь нескольких распределений

Пусть $\xi \sim F(x) = W_1 F_1(x) + \ldots + W_L F_L(x)$

W_k — доля страты $k = \overline{1, L}$

Пусть
$\mu = E\xi$
$\sigma^2 = D\xi$
μ_k — математическое ожидание страты k
σ_k^2 — дисперсия страты k

Мат. ожидание и дисперсия

$$E\xi = \int_{-\infty}^{\infty} x dF(x) = \int_{-\infty}^{\infty} x d(W_1 F_1(x) + \ldots + W_L F_L(x))$$

$$D\xi = \int_{-\infty}^{\infty} (x - E\xi)^2 dF(x) = \int_{-\infty}^{\infty} (x - E\xi)^2 d(W_1 F_1(x) + \ldots + W_L F_L(x))$$

Стратифицированные выборки

В рамках страты k
возьмем выборку объема n_k $(X_{1k}, \ldots, X_{n_k k})$, $k = \overline{1, L}$

Выборочное среднее

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik}$$

Выборочная дисперсия

$$D_k^* = \frac{1}{n_k} \sum_{i=1}^{n_k} (X_{ik} - \bar{X}_k)^2$$

Оценка мат.ожидания смеси

$$\mu = \sum_{k=1}^{L} W_k \mu_k \quad \Rightarrow \quad \bar{X}_S = \sum_{k=1}^{L} W_k \bar{X}_k$$

$$E\bar{X}_S = \sum_{k=1}^{L} W_k E\bar{X}_k = \sum_{k=1}^{L} W_k \mu_k$$

$$D(\bar{X}_S) = \sum_{k=1}^{L} W_k^2 D(\bar{X}_k) = \sum_{k=1}^{L} W_k^2 \frac{\sigma_k^2}{n_k}$$

Оценка дисперсии смеси

$$D\xi = \sum_{k=1}^{L} W_k \sigma_k^2 + \sum_{k=1}^{L} W_k \left(\mu - \mu_k\right)^2 \quad \Rightarrow$$

$$D_S^* = \sum_{k=1}^{L} W_k D_k^* + \sum_{k=1}^{L} W_k \left(\sum_{i=1}^{L} W_i \bar{X}_i - \bar{X}_k \right)^2$$

$$E(D_S^*) = \sum_{k=1}^{L} W_k E D_k^* + \sum_{k=1}^{L} W_k E \left(\sum_{i=1}^{L} W_i \bar{X}_i - \bar{X}_k \right)^2$$

$$D(D_S^*) = \sum_{k=1}^{L} W_k^2 D(D_k^*) + D \left(\sum_{k=1}^{L} W_k \left(\sum_{i=1}^{L} W_i \bar{X}_i - \bar{X}_k \right)^2 \right)$$

Какими выбрать объемы выборок из страт?

Теорема
Объемы выборок $n_1, \ldots, n_L$ такие, что

$$\tilde{n}_k = n \frac{W_k \sigma_k}{\sum_{k=1}^{L} W_k \sigma_k}, \quad k = \overline{1, L}, \quad n = n_1 + \ldots + n_L,$$

минимизируют дисперсию $D(\bar{X}_S)$.

Оптимальное сэмплирование (Неймана): $n_k = \tilde{n}_k, \quad k = \overline{1, L}$

Пропорциональное сэмплирование

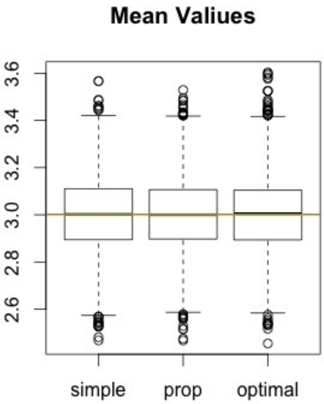
$$n_k = nW_k, \quad k = \overline{1, L}, \quad n = n_1 + \ldots + n_L.$$

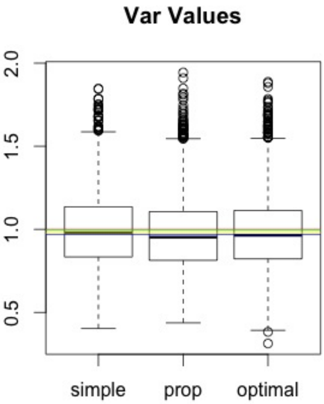
Зачем стратифицировать?

- Уменьшение разброса значений оценок неизвестных параметров.
- Присутствие в выборке представителей всех страт.

РаВные мат.ожидания и раВные дисперсии

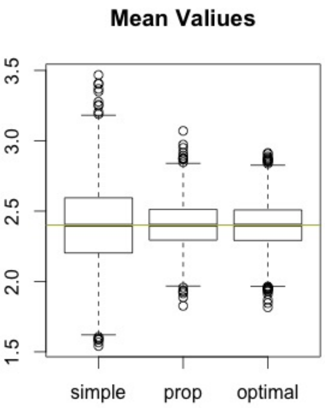
$$E\xi_1 = E\xi_2 = E\xi_3,\ D\xi_1 = D\xi_2 = D\xi_3$$

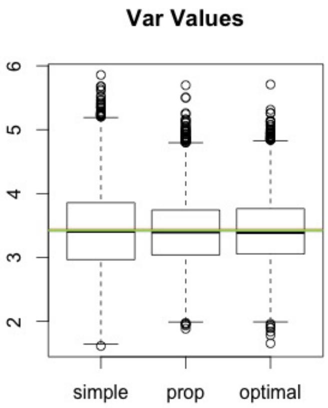




РаЗные мат.ожидания и раВные дисперсии

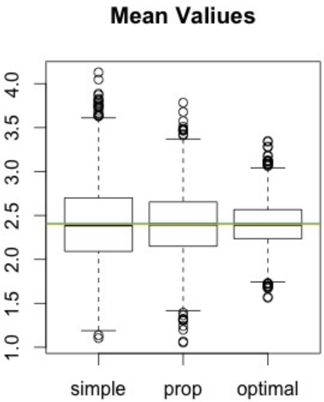
$E\xi_1 \neq E\xi_2 \neq E\xi_3,\ D\xi_1 = D\xi_2 = D\xi_3$

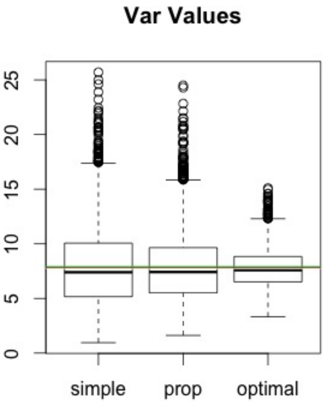




РаЗные мат.ожидания и раЗные дисперсии

$E\xi_1 \neq E\xi_2 \neq E\xi_3$, $D\xi_1 \neq D\xi_2 \neq D\xi_3$





Распределение χ^2

Пусть $\zeta_1,\ldots,\zeta_k \sim N(0,1)$, взаимно независимы

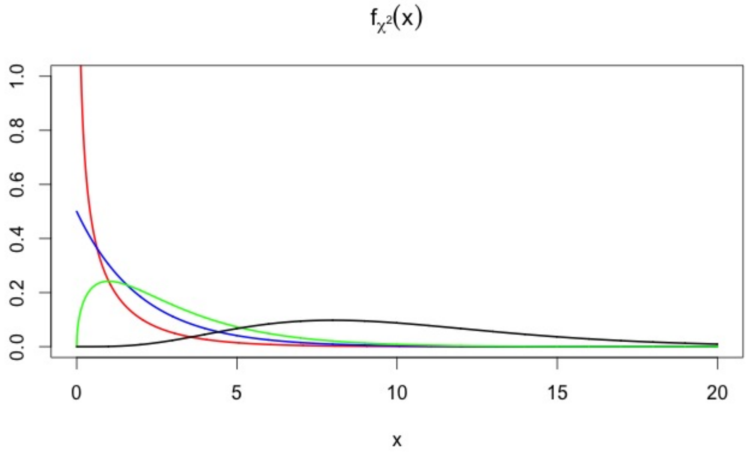
Распределение случайной величины

$$\tau_k = \zeta_1^2 + \ldots + \zeta_k^2$$

называется *распределением хи-квадрат с k степенями свободы*.

$\Gamma(k/2, 1/2)$:

$$f_\tau(x) = \begin{cases} \left(\frac{1}{2}\right)^{\frac{k}{2}} \frac{x^{\frac{k}{2}-1}}{\Gamma(\frac{k}{2})} e^{-\frac{x}{2}}, & x > 0; \\ 0, & x \le 0. \end{cases}$$



$f_{\chi^2}(x)$
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
x

Распределение Стьюдента

Пусть $\zeta \sim N(0,1)$ и $\tau_k \sim \chi^2_k$, взаимно независимы.

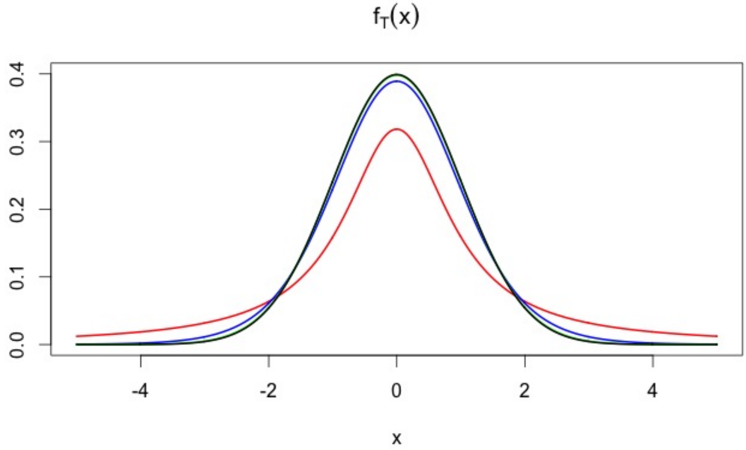
Распределение случайной величины

$$\xi = \frac{\zeta}{\sqrt{\frac{\tau_k}{k}}}$$

называется *распределением Стьюдента с k степенями свободы*.

Плотность распределения T_k:

$$f(z) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k}\Gamma(\frac{k}{2})} \frac{1}{(1 + z^2/k)^{\frac{k+1}{2}}}.$$


$f_T(x)$
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

Распределение Фишера

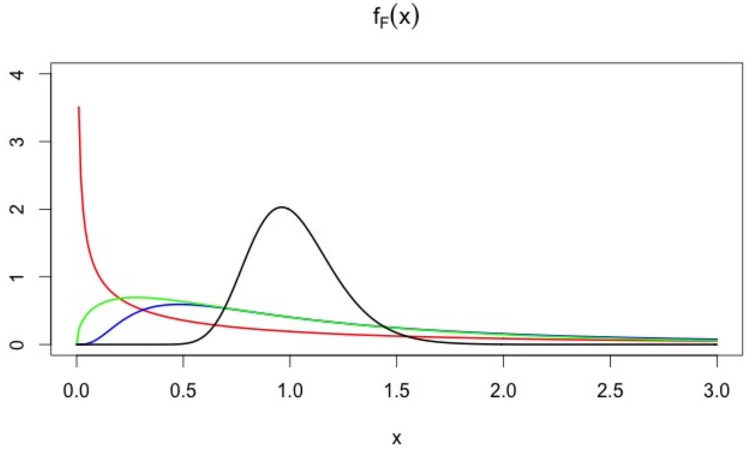
Пусть $\eta \sim \chi^2_m$, $\xi \sim \chi^2_n$ независимы. Будем говорить, что случайная величина

$$\zeta = \frac{\eta/m}{\xi/n}$$

подчиняется *распределению Фишера со степенями свободы числителя m и знаменателя n*.

Плотность распределения ζ:

$$f_\zeta(z) = \begin{cases} \dfrac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \dfrac{m^{\frac{m}{2}} n^{\frac{n}{2}} z^{\frac{m}{2}-1}}{(n+mz)^{\frac{m+n}{2}}}, & \text{если } z > 0; \\ 0, & \text{если } z \leqslant 0. \end{cases}$$


$f_F(x)$
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x