Исследование распределений статистик и мощности критериев однородности в случае больших массивов данных

Выполнил: Федосов Д. Н. , ФПМИ, группа ПММ-61,

Научный руководитель: д.т.н., доцент Чимитова Е.В.

Цель и задачи исследования

Цель исследования: исследование критериев однородности по выборкам большого объема в случае ограниченной точности регистрации наблюдений.

Задачи:

1) Программная реализация вычисления статистик критериев, значений предельных функций распределений соответствующих критериев и оценок мощности критериев.
2) Исследование распределения статистик и мощности критериев.
3) Выявить наиболее предпочтительные критерии.

Критерий Лемана-Розенблатта

$$T = \frac{1}{mn(m+n)}\left[n\sum_{i=1}^{n}\left(r_i - i\right)^2 + m\sum_{j=1}^{m}\left(s_j - j\right)^2 \right] - \frac{4mn-1}{6(m+n)},$$

где r_i – порядковый номер (ранг) y_i ; s_j – порядковый номер (ранг) x_j в объединенном вариационном ряде.

Критерий Лемана-Розенблатта

$$T = \frac{1}{mn(m+n)}\left[n\sum_{i=1}^{n}\left(r_i - i\right)^2 + m\sum_{j=1}^{m}\left(s_j - j\right)^2 \right] - \frac{4mn-1}{6(m+n)},$$

где r_i – порядковый номер (ранг) y_i ; s_j – порядковый номер (ранг) x_j в объединенном вариационном ряде.

$$\lim_{\substack{m\to\infty \\ n\to\infty}} P\left\{T < t\right\} = a1(t),$$

$$a1(t) = \frac{1}{\sqrt{2t}}\sum_{j=0}^{\infty}\frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2}{16t}\right\}\times$$

$$\times\left\{I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16t}\right] - I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16t}\right]\right\},$$

где $I_{-\frac{1}{4}}(*), I_{\frac{1}{4}}(*)$ – модифицированные функции Бесселя.

Критерий Смирнова

$$D_{m,n} = \sup_x \left| G_m(x) - F_n(x) \right|$$

Критерий Смирнова

$$D_{m,n} = \sup_x \left| G_m(x) - F_n(x) \right|$$

$$D^+_{m,n} = \max_{1 \le r \le m} \left[\frac{r}{m} - F_n(x_r) \right] = \max_{1 \le s \le n} \left[G_m(y_s) - \frac{s-1}{n} \right] \qquad D^-_{m,n} = \max_{1 \le r \le m} \left[F_n(x_r) - \frac{r-1}{m} \right] = \max_{1 \le s \le n} \left[\frac{s}{n} - G_m(y_s) \right]$$

$$D_{m,n} = \max\left(D^+_{m,n}, D^-_{m,n} \right)$$

Критерий Смирнова

$$D_{m,n} = \sup_x \left| G_m(x) - F_n(x) \right|$$

$$D^+_{m,n} = \max_{1 \le r \le m} \left[\frac{r}{m} - F_n(x_r) \right] = \max_{1 \le s \le n} \left[G_m(y_s) - \frac{s-1}{n} \right] \qquad D^-_{m,n} = \max_{1 \le r \le m} \left[F_n(x_r) - \frac{r-1}{m} \right] = \max_{1 \le s \le n} \left[\frac{s}{n} - G_m(y_s) \right]$$

$$D_{m,n} = \max\left(D^+_{m,n}, D^-_{m,n} \right)$$

$$S_C = \sqrt{\frac{mn}{m+n}} D_{m,n}$$

$$\lim_{\substack{m \to \infty \\ n \to \infty}} P\left\{ S_C < S \right\} = K(S), \quad K(s) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}$$

Критерий Андерсона-Дарлинга

$$A^2 = \frac{1}{mn} \sum_{i=1}^{m+n-1} \frac{\left(M_i(m+n) - mi\right)^2}{i(m+n-i)},$$

где M_i – число элементов первой выборки, меньших или равных i–му элементу вариационного ряда объединенной выборки.

Критерий Андерсона-Дарлинга

$$A^2 = \frac{1}{mn}\sum_{i=1}^{m+n-1}\frac{\left(M_i(m+n)-mi\right)^2}{i(m+n-i)},$$

где M_i – число элементов первой выборки, меньших или равных i–му элементу вариационного ряда объединенной выборки.

$$a2(t) = \frac{\sqrt{2\pi}}{t}\sum_{j=0}^{\infty}(-1)^j\frac{\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2\pi^2}{8t}\right\}\times$$
$$\times\int_0^{\infty}\exp\left\{\frac{t}{8(y^2+1)}-\frac{(4j+1)^2\pi^2y^2}{8t}\right\}dy$$

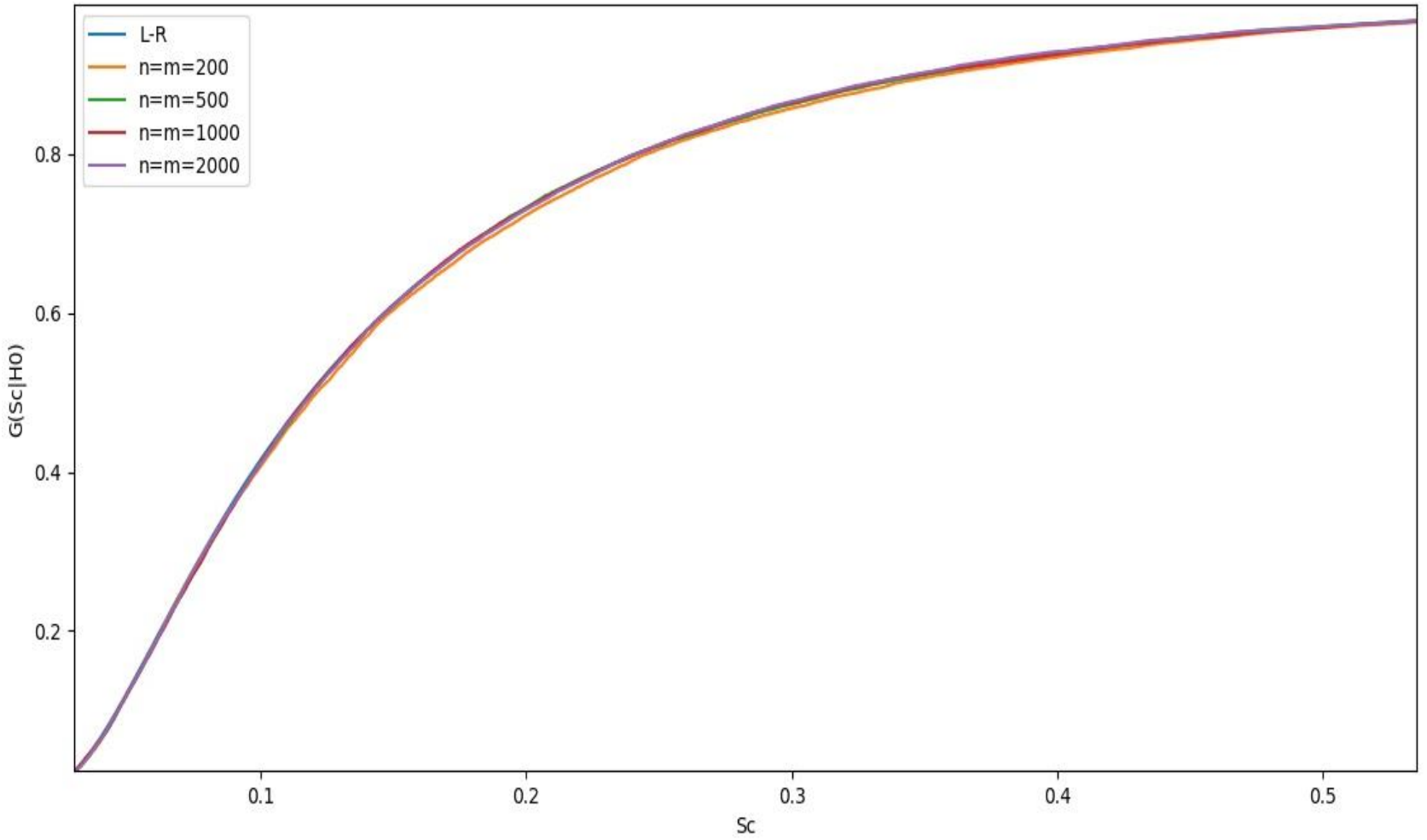
Исследование распределений статистик

- Объем моделирования $N = 16600$

- $\rho = \sup_x |F_n(x) - F(x)|$ - расстояние между эмпирической и предельной функциями распределения статистик критерия в метрике Колмогорова. Где $F_n(x)$ - эмпирическая функция распределения по вычисленным значениям статистик; $F(x)$ – предельная функция распределения статистики критерия.

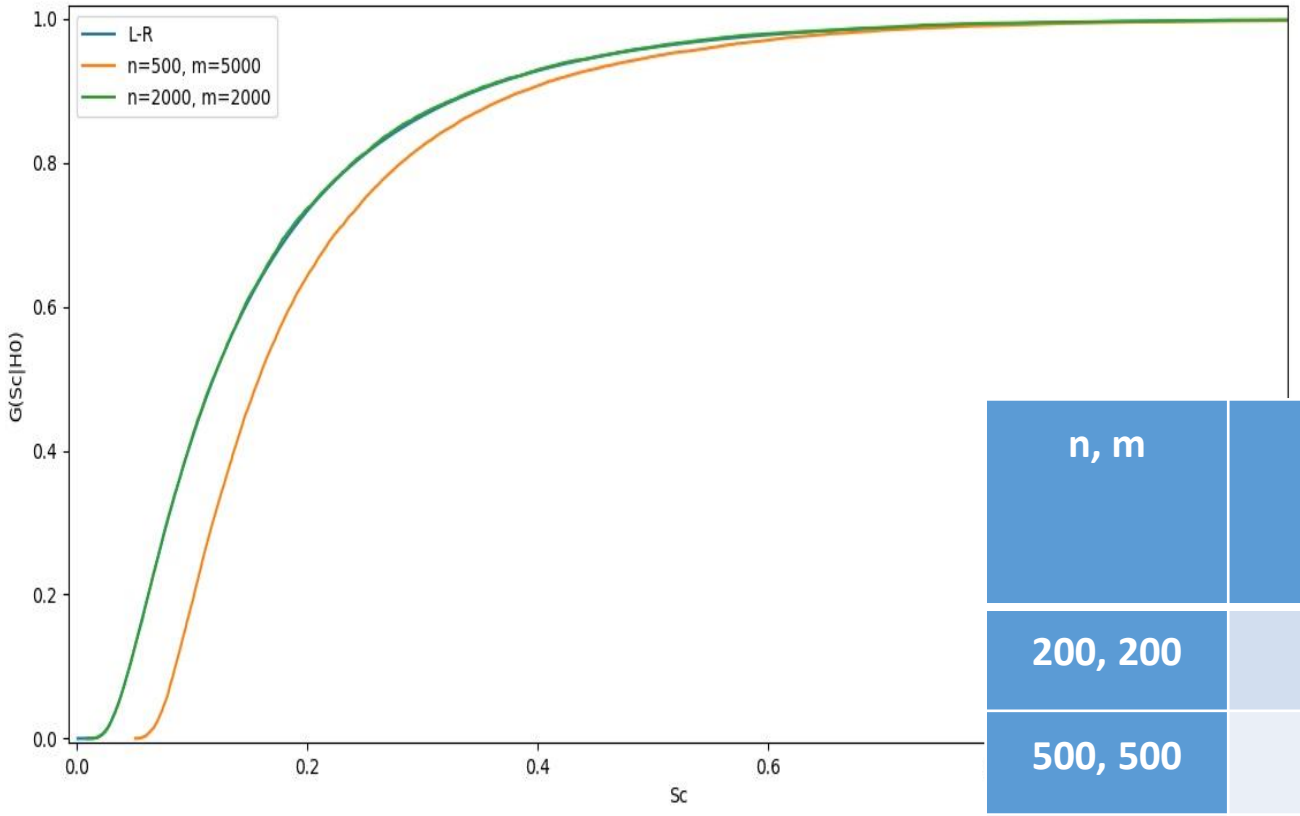
Исследование распределений статистик критерия Лемана-Розенблатта

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.01	243.5
500, 500	0.01	369.5
1000, 1000	0.01	448.5
2000, 2000	0.01	510.5
5000, 5000	0.01	578.5



Округление до 2 знаков, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$

Исследование распределений статистик критерия Лемана-Розенблатта, $n \neq m$



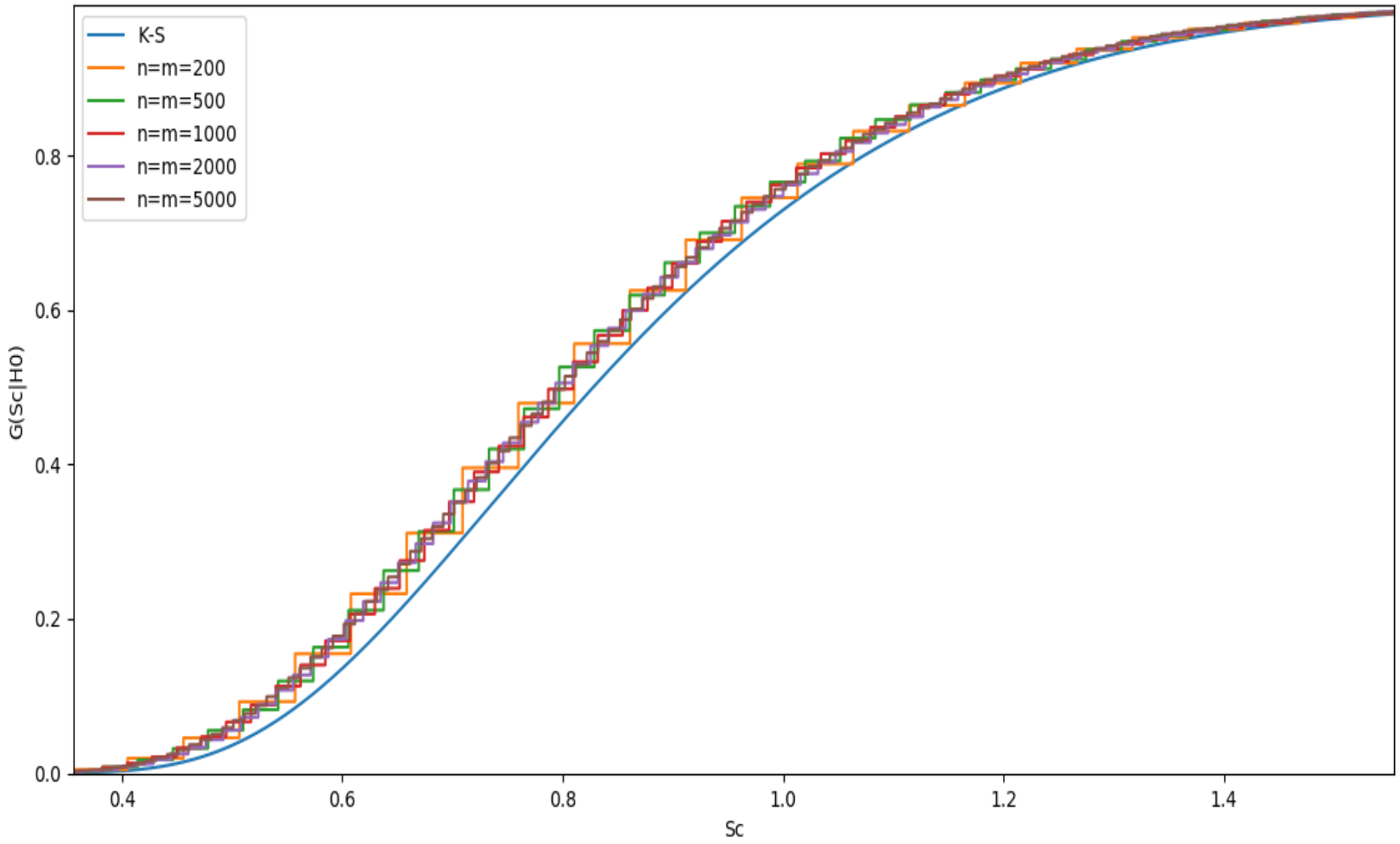
n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.01	243.5
500, 500	0.01	369.5
1000, 1000	0.01	448.5
2000, 2000	0.01	510.5
5000, 5000	0.01	578.5

n, m	ρ	среднее число различных значений в объединенной выборке
500, 500	0.01	369.5
500, 1000	0.01	418.0
500, 2000	0.03	469.0
500, 5000	0.24	535

Исследование распределений статистик критерия Смирнова

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.09	246.5
500, 500	0.07	368.5
1000, 1000	0.07	449.0
2000, 2000	0.07	507.0
5000, 5000	0.06	580.5

Округление до 2 знаков, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$



Исследование распределений статистик критерия Смирнова, $n \neq m$

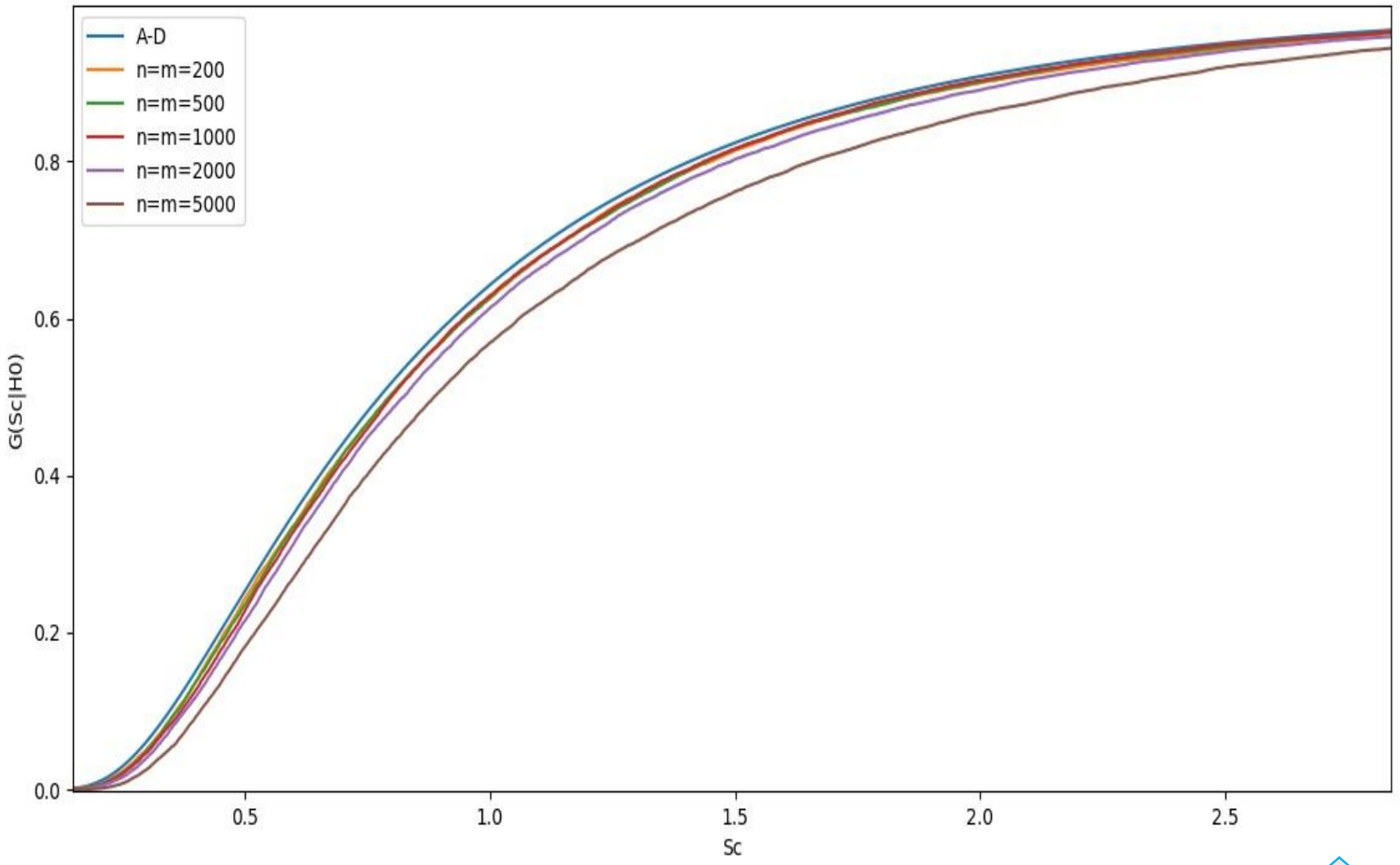
n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.09	246.5
500, 500	0.07	368.5
1000, 1000	0.07	449.0
2000, 2000	0.07	507.0
5000, 5000	0.06	580.5

n, m	ρ	среднее число различных значений в объединенной выборке
500, 500	0.07	368.5
500, 1000	0.07	421.05
500, 2000	0.06	468.0
500, 5000	0.05	533.5

Исследование распределений статистик критерия Андерсона-Дарлинга

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.02	241.0
500, 500	0.02	377.0
1000, 1000	0.03	442.0
2000, 2000	0.04	510.0
5000, 5000	0.08	576.5

Округление до 2 знаков, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$



Исследование распределений статистик критерия Андерсона-Дарлинга

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.02	241.0
500, 500	0.02	377.0
1000, 1000	0.03	442.0
2000, 2000	0.04	510.0
5000, 5000	0.08	576.5

Округление до 2 знаков, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$

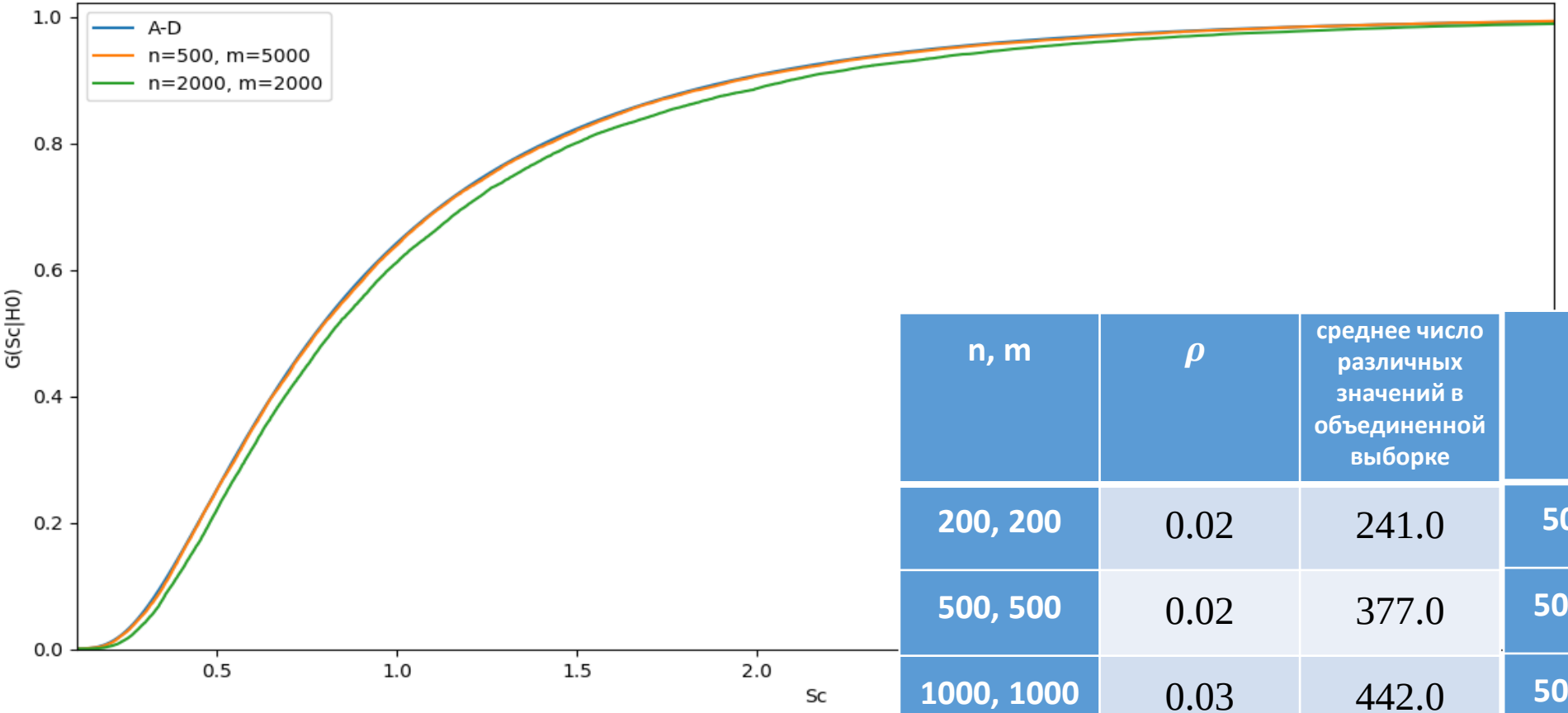
n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.02	249.0
500, 500	0.02	374.5
1000, 1000	0.02	442.5
2000, 2000	0.04	503.5
5000, 5000	0.09	569.0

Округление до 1 знака, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 10$

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.02	221.0
500, 500	0.03	321.0
1000, 1000	0.04	374.0
2000, 2000	0.06	421.5
5000, 5000	0.12	475.0

Округление до целых, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 80$

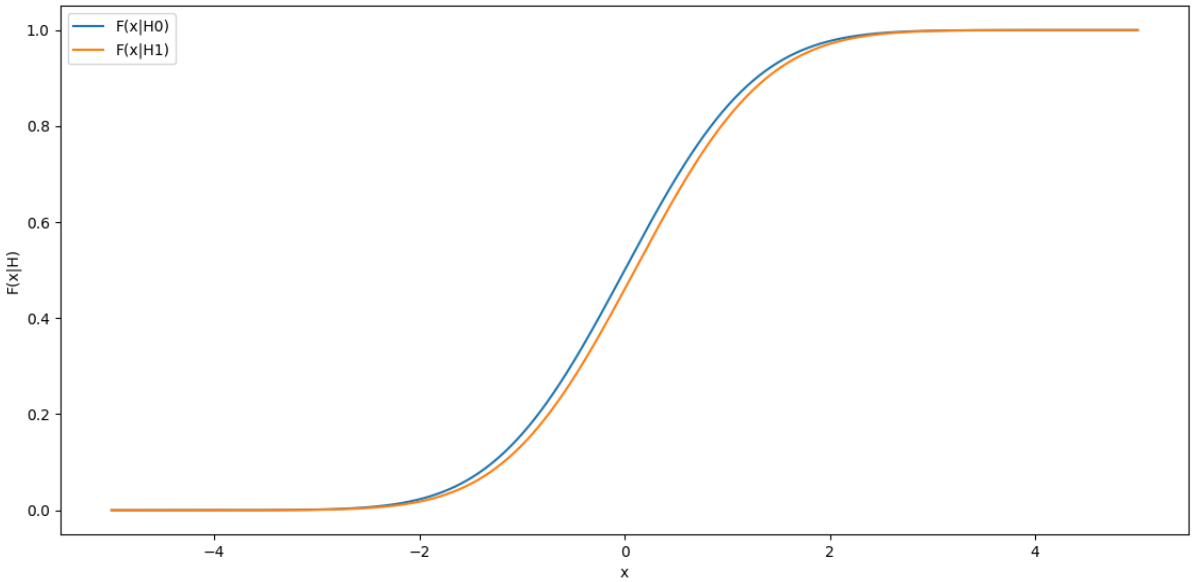
Исследование распределений статистик критерия Андерсона-Дарлинга, $n \neq m$

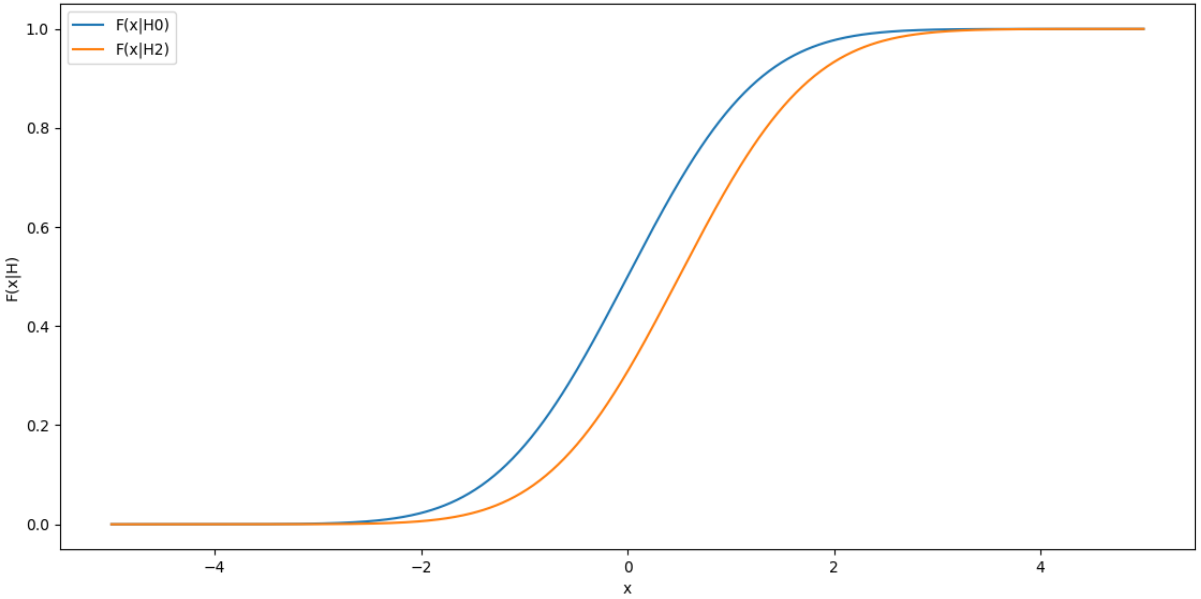


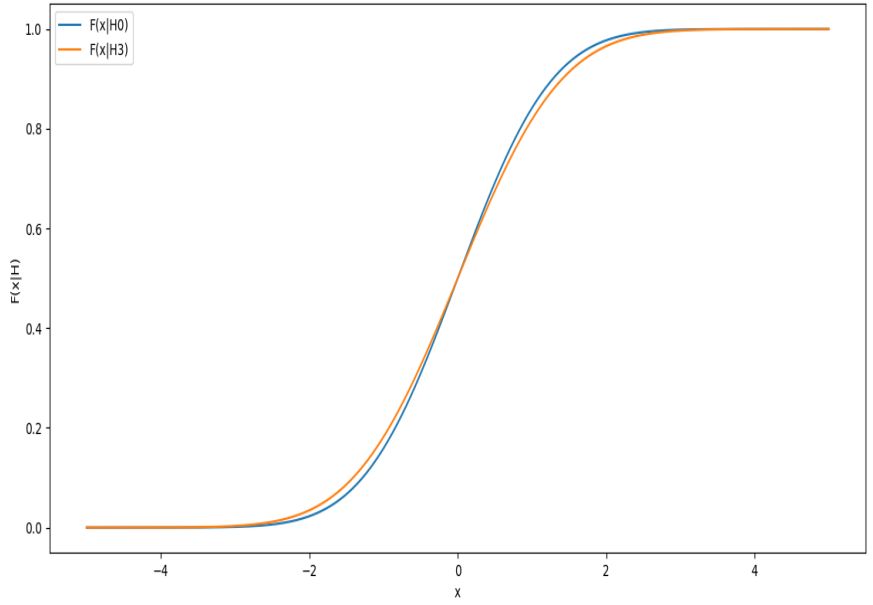
n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.02	241.0
500, 500	0.02	377.0
1000, 1000	0.03	442.0
2000, 2000	0.04	510.0
5000, 5000	0.08	576.5

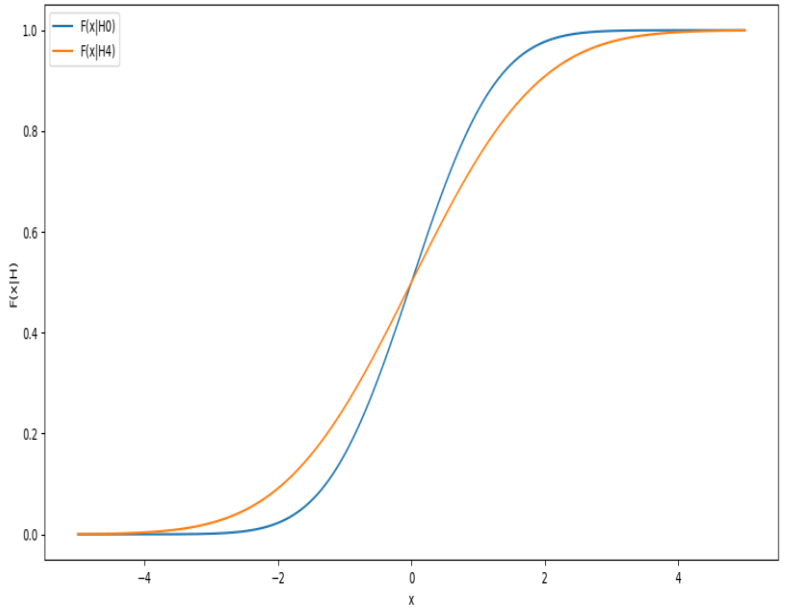
n, m	ρ	среднее число различных значений в объединенной выборке
500, 500	0.02	377.0
500, 1000	0.01	422.0
500, 2000	0.01	465.0
500, 5000	0.01	532.5

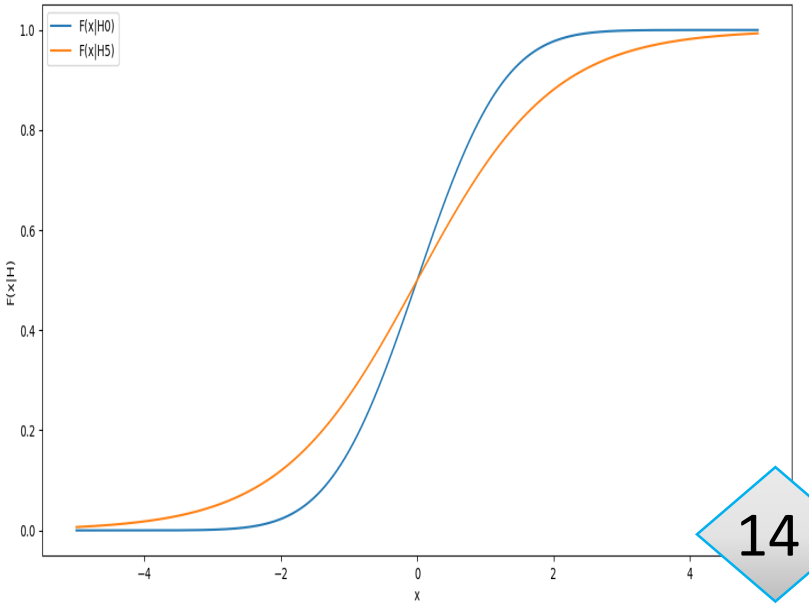
Виды распределения в альтернативах относительно стандартного нормального распределения



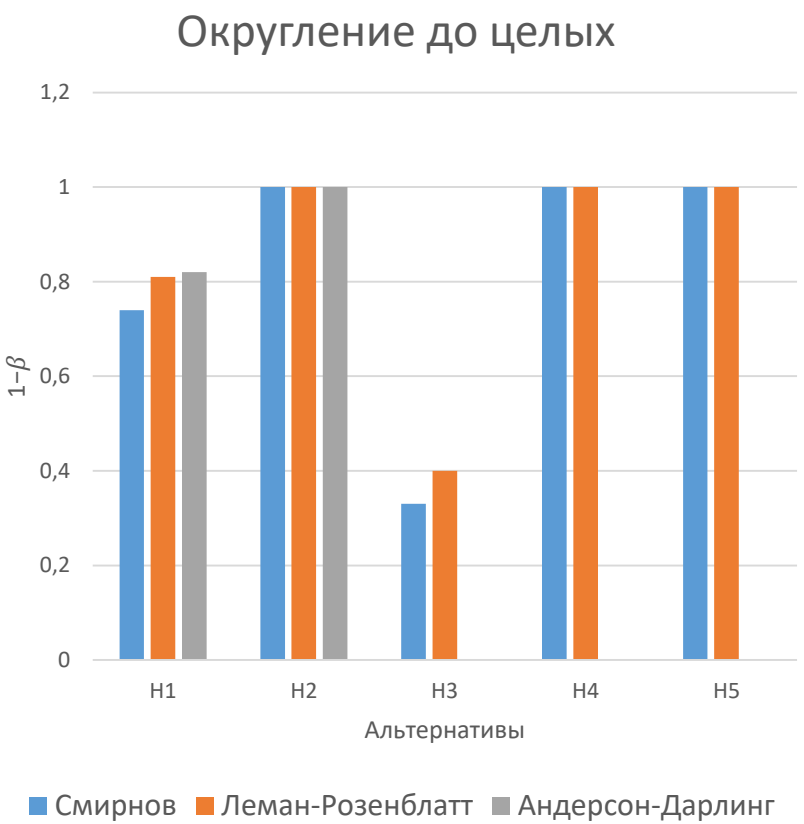


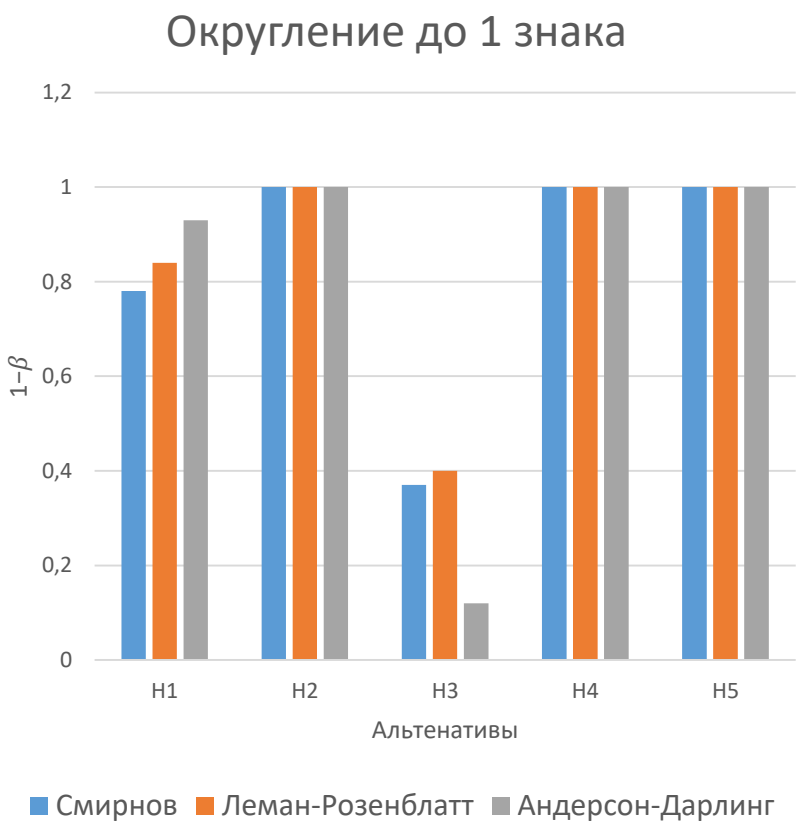


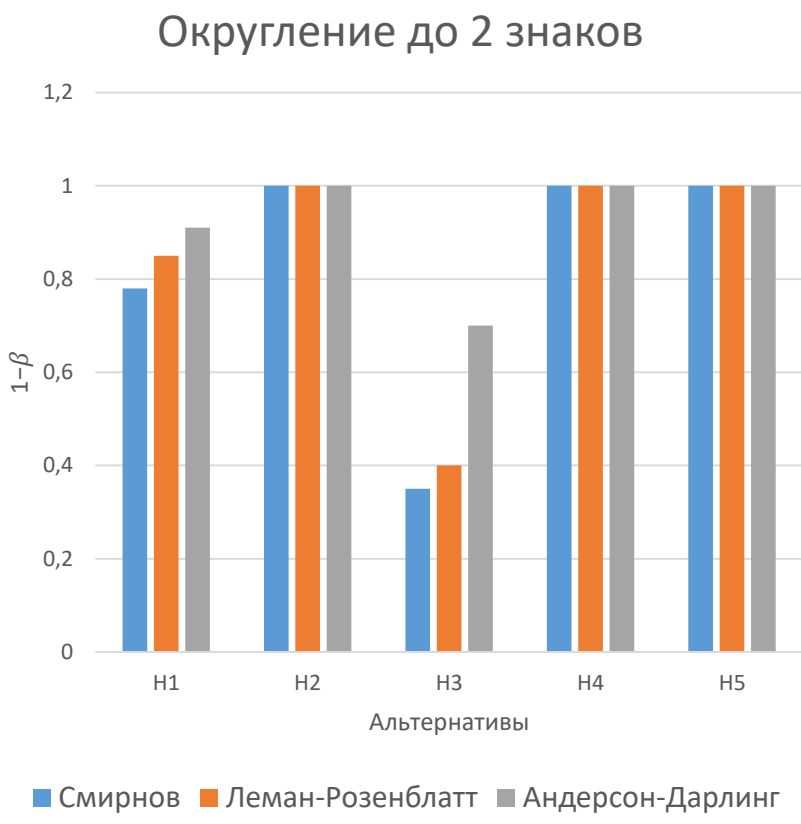




Сравнительный анализ мощности критериев, n=m=2000, α=0.05







Заключение

- ✓ для критерия Лемана-Розенблатта распределения статистики остаются близкими к предельному закону при равных объемах выборок, однако при $n \neq m$ расстояние между эмпирической функцией распределения статистики и предельным увеличивается с ростом разницы в объемах выборок;
- ✓ для критерия Смирнова наблюдается довольно медленная сходимость распределения статистики к предельному закону при увеличении объемов выборок при $n = m$. Но при $n \neq m$, расстояние между функцией распределения статистики и предельной функцией становится несколько меньше, чем в случае $n = m$;
- ✓ для критерия Андерсона-Дарлинга расстояние между эмпирической функцией распределения статистики и предельным уменьшается с ростом отношения числа различных значений в объединенной выборке к объему объединенной выборки при $n = m$; А также, наблюдается сближение распределение статистики к предельному распределению при увеличении разности объемов выборок;

Заключение

- ✓на данных ограниченной точности наибольшую мощность среди рассмотренных критериев показали критерии Андерсона-Дарлинга и Лемана-Розенблатта. Однако в случае округления наблюдений в выборках до целых критерий Андерсона-Дарлинга оказался смещенным относительно конкурирующих гипотез с пересечением функций распределения;
- ✓обобщая полученные результаты, можно сделать вывод о предпочтительности использования критерия Лемана-Розенблатта при равных объемах выборок $n = m$.