О.А.Демидова

Лекция Непараметрические методы в статистике

Тест Колмогорова-Смирнова о равенстве распределений

Нередко при решении практических экономических задач перед исследователем встает вопрос: можно ли считать, что две выборки $X_1,...,X_m$ и $Y_1,...,Y_n$ принадлежат одной генеральной совокупности? Иногда представляет интерес проверка гипотезы не о полном совпадении распределений, а только их некоторых числовых характеристик, например, математических ожиданий или медиан.

Если делается дополнительное предположение о нормальном распределении генеральной совокупности или выборки достаточно большого размера, то при проверке гипотез о равенстве средних используются t-тесты.

Однако если перечисленные выше предположения не выполняются, то «на помощь приходят» непараметрические тесты.

Наиболее общим из таких тестов является тест Колмогорова – Смирнова, с помощью которого проверяется гипотеза о совпадении распределений.

Этот тест может быть проведен для двух выборок (в этом случае тестируется гипотеза о совпадении соответствующих распределений без фиксации заранее семейства распределений, к которому принадлежат эти выборки) или для одной выборки (в этом случае семейство распределений, принадлежность к которому проверяется, задается заранее).

Начнем с описания теста Колмогорова-Смирнова для двух выборок. Пусть $X_1,...,X_m$ - выборка из генеральной совокупности с функцией распределения F, а $Y_1,...,Y_n$ - выборка из генеральной совокупности с функцией распределения G, причем эти выборки являются взаимно независимыми, а функции F и G заранее неизвестны.

Тогда нулевая гипотеза имеет следующий вид:

$H_0 : F(t) = G(t)$ для любого t,

А альтернативная гипотеза:

$H_1 : F(t) \neq G(t)$ при некотором t.

Для проверки этой гипотезы используется статистика J Колмогорова - Смирнова

$$J = \frac{mn}{d} \max_{-\infty<t<\infty} |F_m(t) - G_n(t)|,$$

Где $F_m(t)$, $G_n(t)$ - эмпирические функции распределения для выборок $X_1,...,X_m$ и $Y_1,...,Y_n$, а именно,

$$F_m(t) = \frac{\textit{число элементов в первой выборке, не превышающих } t}{m},$$

$$G_n(t) = \frac{\textit{число элементов во второй выборке, не превышающих } t}{n},$$

d – наибольший общий делитель m и n.

Обе эмпирические функции являются кусочно-непрерывными со «скачками» в значениях $X_1,...,X_m$ и $Y_1,...,Y_n$ соответственно.

Объединив и упорядочив значения двух выборок $X_1,...,X_m$ и $Y_1,...,Y_n$, положив $Z_{(1)} \le ... \le Z_{(N)}$, где N = m + n, мы можем переписать тестовую статистику в более удобном для вычислений виде:

$$J = \frac{mn}{d} \max_{i=1,...,N} \left| F_m(Z_{(i)}) - G_n(Z_{(i)}) \right|$$

Замечательное свойство статистики Колмогорова – Смирнова J состоит в том, что ее распределение не зависит от конкретного вида F и G. Критические значения для статистики Колмогорова-Смирнова ($j_\alpha(m,n)$ для соответствующего уровня значимости α) могут быть найдены в таблице П1 приложения 1. Правило выбора между основной и альтернативной гипотезами имеет следующий вид: при уровне значимости α гипотеза H_0 отвергается в пользу гипотезы H_1 при выполнении условия $J \ge j_\alpha$ (а в противном случае H_0 не отвергается).

Если $\min(m,n) \to \infty$, т.е. в случае больших выборок, может быть использовано предельное распределение статистики

$$J^* = \left(\frac{mn}{N} \right)^{1/2} \max_{i=1,...,N} \left| F_m(Z_{(i)}) - G_n(Z_{(i)}) \right| = \frac{d}{(mnN)^{1/2}} J ,$$

А именно,

$$P_0(J^* < s) \to \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2} \quad для\ s > 0$$ и 0 в противном случае.

В этом случае нет необходимости использовать достаточно громоздкие таблицы и правило выбора между основной и альтернативной гипотезами значительно упрощается.

При больших выборках гипотеза H_0 отвергается в пользу гипотезы H_1 при выполнении условия:

$J^* \ge q_\alpha^*$; где q_α^* определяется из соотношения $Q(q_\alpha^*) = \alpha$, а

$$Q(s) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2} \quad для\ \ s > 0$$ (и 0 в противном случае).

При применении теста Колмогорова-Смирнова для проверки, что некоторая выборка принадлежит конкретному семейству распределений с известной функцией распределения $F_0(t)$:

$H_0 : F(t) = F_0(t)$ для любого t

При альтернативной гипотезе

$H_1 : F(t) \ne F_0(t)$, при некотором t.

Статистика Колмогорова-Смирнова редуцируется к виду:

$$J_0 = \max_{-\infty < t < \infty} \left| F_m(t) - F_0(t) \right| ,$$

Или к более удобной для вычислений форме:

$$J_0 = \max_{i=1,...,m} \left| F_m(X_i) - F_0(X_i) \right|$$

.

Критические значения для статистики J_0 могут быть найдены в таблице П2 приложения 1.

Однако, как и в предыдущем случае, на практике чаще используют предельные распределения. При $m \to \infty$ используется следующая аппроксимация:

$P_0(\sqrt{m}J_0 < s) \to 1 - 2\sum_{k=0}^{\infty}(-1)^k e^{-2k^2s^2}$ *при* $s > 0$ и 0 в противном случае.

В следующих разделах будут приведены тесты для проверки более «мягких» гипотез о совпадении не полном совпадении распределений для двух выборок, а их средних (математических ожиданий) или медиан.

Тест Уилкоксона – Манна - Уитни для проверки гипотезы о равенстве средних для двух выборок

Как уже отмечалось выше, в некоторых случаях нас интересует проверка гипотезы не о полном совпадении распределений для двух рассматриваемых выборок, а, например, о равенстве их математических ожиданий. В этом случае применяется описанный ниже критерий суммы рангов Уилкоксона-Манна-Уитни.

Пусть у нас опять имеются две независимые выборки $X_1,...,X_m$ ($E(X_1) = ... = E(X_m) = E(X)$) и $Y_1,...,Y_n$ ($E(Y_1) = ...E(Y_n) = E(Y)$). Обозначим $\Delta = E(X) - E(Y)$.

Тогда гипотезу о равенстве математических ожиданий можно сформулировать следующим образом:

$H_0 : \Delta = 0$

Альтернативная гипотеза может быть

двусторонней: 1) $H_1 : \Delta \neq 0$ (т.е. математические ожидания для двух генеральных совокупностей различаются)

или односторонней, причем возможны два варианта:

2) $H_1 : \Delta > 0$ (т.е. математическое ожидание первой генеральной совокупности больше),

3) $H_1 : \Delta < 0$ (т.е. математическое ожидание первой генеральной совокупности меньше),

Для вычисления тестовой статистики W элементы обеих выборок объединяются в общую выборку длины N = m + n и вычисляется ранг $S_1,...,S_N$ каждого элемента. Напомним, что ранг элемента – это номер, который он получает в общей упорядоченной выборке, если все элементы в обеих выборках различны. Если же среди элементов объединенной выборки есть совпадающие, то им приписывают одинаковые связанные ранги. Каждый из связанных элементов получает ранг, равный среднему арифметическому рангов, который имели бы элементы связки, если бы различались.

Тестовая статистика Уилкоксона-Манна-Уитни имеет вид:

$W = \sum_{j=1}^{n} S_j$, т.е. суммируются ранги только одной выборки.

Критические значения для статистики W (при заданных уровнях значимости) w_α могут быть найдены в таблице П3 приложения 1.

Гипотеза H_0 при выбранном уровне значимости α отвергается в пользу гипотезы

1) $H_1 : \Delta \neq 0$

если $W \geq w_{\alpha/2}$ или $W \leq n(m+n+1) - w_{\alpha/2}$,

в пользу гипотезы

2) $H_1 : \Delta > 0$

если $W \geq w_\alpha$,

в пользу гипотезы

3) $H_1 : \Delta < 0$

если $W \leq n(m+n+1) - w_\alpha$.

Однако на практике и в этом случае чаще используют предельное распределение, используя тот факт, что при $\min(m,n) \to \infty$ (случай больших выборок), распределение статистики W является асимптотически нормальным

С математическим ожиданием $E_0(W) = \frac{n(m+n+1)}{2}$,

И дисперсией $\mathrm{var}_0(W) = \frac{mn(m+n+1)}{12}$ (в случае отсутствия связанных рангов).

Центрируя и нормируя W: $W^* = \frac{W - E_0(W)}{\sqrt{\mathrm{var}_0(W)}}$,

приходим к стандартному нормальному распределению:

$W^* \sim N(0,1)$.

Правило выбора между основной и альтернативной гипотезой принимает вид:

Гипотеза H_0 при выбранном уровне значимости α отвергается в пользу гипотезы

1) $H_1 : \Delta \neq 0$

если $|W^*| \geq z_{\alpha/2}$,

в пользу гипотезы

2) $H_1 : \Delta > 0$

если $W^* \geq z_\alpha$,

в пользу гипотезы

3) $H_1 : \Delta < 0$

если $W^* \leq -z_\alpha$.

В случае связанных рангов формула для математического ожидания не изменяется, а для дисперсии усложняется:

$$\mathrm{var}_0(W) = \frac{mn}{12}\left[m+n+1 - \frac{\sum_{j=1}^{g}(t_i-1)(t_j+1)}{(m+n)(m+n-1)} \right],$$

или $\mathrm{var}_0(W) = \frac{mn(N+1)}{12} - \frac{mn}{12N(N-1)} \cdot \sum_{j=1}^{g}(t_j-1)t_j(t_j+1)$,

где g обозначает число групп связанных элементов, $t_j, j = 1,...,g$ - размер связанной j-ой группы.

Тесты для проверки гипотез о равенстве медиан для двух выборок

В некоторых случаях необходимо сравнить не математические ожидания, а медианы генеральных совокупностей, выборки из которых мы рассматриваем. Приведем соответствующие тесты.

Начнем с наиболее простого случая, когда наблюдения из двух выборок сгруппированы парами $(X_i, Y_i), i = 1,...,n$.

Обозначим $Z_i = Y_i - X_i$ для $i = 1,...,n$ и предположим, что каждая Z_i, $i = 1,...,n$ является реализацией случайной величины с функцией распределения F_i, $i = 1,...,n$ (для различных i эти функции могут не совпадать), причем $F_i(\theta + t) + F_i(\theta - t) = 1$, для всех t и $i = 1,...,n$, что соответствует симметричности всех распределений относительно общей медианы θ. Этот параметр обычно интерпретируется как эффект воздействия.

Гипотеза о равенстве медиан для двух распределений интерпретируется как отсутствие эффекта воздействия:

$H_0 : \theta = 0$.

Альтернативная гипотеза, как и в предыдущем случае, может быть

двусторонней $H_1 : \theta \neq 0$,

либо односторонней (возможны два варианта)

$H_1 : \theta > 0$ или 3) $H_1 : \theta < 0$

Для проверки этой гипотезы существует два непараметрических теста (Уилкоксона и Фишера).

При проведении теста Уилкоксона упорядочиваются модули разностей $|Z_1|,...,|Z_n|$ и вычисляются их ранги R_i, $i = 1,...,n$. Также определяются индикаторы разностей:

$$\psi_i = \begin{cases} 1, & if\ Z_i > 0, \\ 0, & if\ Z_i < 0 \end{cases} \quad i = 1,...,n.$$

Тестовая статистика Уилкоксона вычисляется как сумма положительных рангов: $T^+ = \sum_{i=1}^{n} R_i \psi_i$.

Критические значения для статистики T^+ (при заданных уровнях значимости) t_α могут быть найдены в таблице П3 приложения 1.

Гипотеза H_0 при выбранном уровне значимости α отвергается в пользу гипотезы

1) $H_1 : \theta \neq 0$

если $T^+ \geq t_{\alpha/2}$ или $T^+ \leq \frac{n(n+1)}{2} - t_{\alpha/2}$

в пользу гипотезы

2) $H_1 : \theta > 0$

если $T^+ \geq t_{\alpha}$,

в пользу гипотезы

3) $H_1 : \theta < 0$

если $T^+ \leq \frac{n(n+1)}{2} - t_{\alpha}$.

Однако на практике и в этом случае чаще используют предельное распределение,

Как и в предыдущих случаях, популярно использование предельного распределения, в данном случае нормального

С математическим ожиданием

$E_0(\mathrm{T}^+) = \frac{n(n+1)}{4}$,

И дисперсией $\mathrm{var}_0(T^+) = \frac{n(n+1)(2n+1)}{24}$ (в случае отсутствия связанных рангов)..

Как и ранее, центрируя и нормируя T^+ : $T* = \frac{T^+ - E_0(T^+)}{\sqrt{\mathrm{var}(T^+)}}$, приходим к стандартному нормальному распределению и применяем традиционное правило для выбора между гипотезами.

Гипотеза H_0 при выбранном уровне значимости α отвергается в пользу гипотезы

1) $H_1 : \theta \neq 0$,

если $|T*| \geq z_{\alpha/2}$,

в пользу гипотезы

2) $H_1 : \theta > 0$,

если $T* \geq z_{\alpha}$,

в пользу гипотезы

3) $H_1 : \theta < 0$,

если $T* \leq -z_{\alpha}$.

Как и в ранее рассмотренном случае, при наличии связанных рангов формула для математического ожидания не изменяется, а для дисперсии усложняется:

$$\mathrm{var}_0(T^+) = \frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_{j=1}^{g} t_j(t_j-1)(t_j+1) \right],$$

где, как и ранее, g обозначает число групп связанных элементов, $t_j, j = 1,...,g$ - размер связанной j-ой группы.

В тесте знаков Фишера тестовой статистикой является сумма (или же количество) положительных разностей: $B = \sum_{i=1}^{n_1} \psi_i$ (отметим, что из выборки $Z_1,...,Z_n$ предварительно удаляются нулевые элементы, n_1 – количество ненулевых элементов).

Поскольку ψ_i, $i = 1,...,n$ принимают значения 1 и 0 (с вероятностью 0.5), то тестовая статистика имеет биномиальное распределение с параметром 0.5.

Критические значения для биномиального распределения (при заданных уровнях значимости) b_α могут быть найдены в таблице П4 приложения 1.

Алгоритм схемы выбора между гипотезами традиционен.

Гипотеза H_0 при выбранном уровне значимости α отвергается в пользу гипотезы

1) $H_1 : \theta \neq 0$,

если $B \geq b_{\alpha/2,1/2}$ или $B \leq n_1 - b_{\alpha/2,1/2}$,

в пользу гипотезы

2) $H_1 : \theta > 0$,

если $B \geq b_{\alpha,1/2}$,

в пользу гипотезы

3) $H_1 : \theta < 0$,

если $B \leq n_1 - b_{\alpha,1/2}$.

При большой выборке, как и в предыдущих случаях, обычно используют предельное распределение (оно будет нормальным).

Учитывая, что при p = 0.5,

$$E_0(\mathrm{B}) = \frac{n}{2},\ \mathrm{var}_0(B) = \frac{n}{4},$$

Традиционная операция центрирования и нормирования B:

$$B^* = \frac{B - E_0(B)}{\sqrt{\mathrm{var}(B)}} = \frac{B - \frac{n}{2}}{\sqrt{n/4}}$$

Приводит к стандартному нормальному распределению N(0,1) и аналогичной использованной ранее схеме выбора между гипотезами:

Гипотеза H_0 при выбранном уровне значимости α отвергается в пользу гипотезы

1) $H_1 : \theta \neq 0$,

если $|B^*| \geq z_{\alpha/2}$,

в пользу гипотезы

2) $H_1 : \theta > 0$,

если $B^* \geq z_\alpha$,

в пользу гипотезы

3) $H_1 : \theta < 0$,

если $B^* \leq -z_\alpha$.

Более общий случай проверки гипотез о совпадении медиан будет рассмотрен в следующем разделе. Количество выборок при этом не ограничивается двумя и они могут иметь разный размер.

Тест Крускалла-Уоллеса для проверки гипотезы о равенстве медиан нескольких выборок

Предположим, у нас имеется k выборок из нескольких генеральных совокупностей, n_j - количество элементов в j-ой выборке, $j = 1,...,k$, F_j - функция распределения генеральной совокупности в j-ой выборки $\{X_{1j}, X_{2j},..., X_{n_j j}\}$.

Мы предполагаем, что N ($N = \sum_{j=1}^{k} n_j$) случайных величин $\{X_{1j}, X_{2j},..., X_{n_j j}\}$, $j = 1,...,k$, в объединенной выборке взаимно независимы.

Для того, чтобы иметь возможность проверить гипотезу о равенстве медиан всех генеральных совокупностей, мы предполагаем, что функции распределения $F_1,...,F_k$ связаны между собой следующим соотношением: $F_j(t) = F(t - \tau_j)$, для $j = 1,...,k$, где F – непрерывная функция распределения с неизвестной медианой θ, τ_j, $j = 1,...,k$ - эффект воздействия для j – ой выборки.

Тогда проверка гипотезы о равенстве медиан для всех выборок будет равносильна проверке гипотезы о равенстве эффектов воздействия $\tau_1,...,\tau_k$:

$H_0 : \tau_1 = ... = \tau_k$

При альтернативной гипотезе

$H_1 : \tau_1,...,\tau_k$ не все равны между собой.

Для вычисления тестовой статистики сначала необходимо объединить все наблюдения в одну общую выборку, упорядочить их и вычислить ранг каждого элемента r_{ij}, где $j = 1,...,k$ - номер выборки, $i = 1,...,n_j$ - номер элемента (X_{ij}) в j – ой выборке. Затем необходимо найти сумму рангов в каждой выборке: $R_j = \sum_{i=1}^{n_j} r_{ij}$

и средний ранг для каждой выборки $R_{\cdot j} = \frac{R_j}{n_j}$, $j = 1,...,k$.

При отсутствии связанных рангов тестовая статистика Краскела-Уоллиса H может быть вычислена по следующей формуле:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(R_{\cdot j} - \frac{N+1}{2} \right)^2 = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1).$$

Отметим, что $\frac{N+1}{2} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{r_{ij}}{N}$ - средний ранг по объединенной выборке.

Критические значения для статистики Краскела-Уоллиса (при заданных уровнях значимости) h_α могут быть найдены в таблице П5 приложения 1.

Правило принятия решения при выборе между гипотезами аналогично приведенным выше.

Гипотеза $H_0 : \tau_1 = ... = \tau_k$ при выбранном уровне значимости α отвергается в пользу альтернативной, если выполняется условие $H \geq h_\alpha$.

При наличии связанных рангов формула для тестовой статистики несколько изменяется и принимает следующий вид:

$$H' = \frac{H}{1 - \left[\sum_{j=1}^{g} (t_j^3 - t_j)/(N^3 - N) \right]},$$

где, как и в предыдущих случаях, g обозначает число групп связанных элементов, $t_j, j = 1,...,g$ - размер связанной j-ой группы.

Если $\min(n_1,...,n_k) \to \infty$ (т.е. рассматриваемые выборки велики), то в качестве аппроксимации используют предельное распределение, в данном случае распределение «хи-квадрат» с $k-1$ степенями свободы.

Критические значения для распределения «хи-квадрат» (при заданных уровнях значимости) могут быть найдены в таблице П6 приложения 1.

Основная гипотеза H_0 при выбранном уровне значимости α отвергается в пользу альтернативной, если выполняется условие $H \geq \chi^2_{k-1,\alpha}$.

Проведение непараметрических тестов с помощью пакетов STATA и R.

Для проведения теста Колмогорова-Смирнова на равенство распределений в статистическом пакете STATA используется команда

ksirnov varname, by (group),

где varname – имя переменной, объединяющей наблюдения из первой и второй выборки,

group – переменная, принимающая два значения, соответствующие принадлежности наблюдения к первой или второй выборке.

Для проведения теста Уилкоксона – Манна - Уитни гипотезы о равенстве средних для двух выборок в статистическом пакете STATA используется команда ranksum varname, by (groupvar),

где varname – имя переменной, объединяющей наблюдения из первой и второй выборки,

groupvar – переменная, принимающая два значения, соответствующие принадлежности наблюдения к первой или второй выборке.

Для проведения теста Уилкоксона о равенстве медиан для двух выборок в статистическом пакете STATA используется команда

signrank X=Y

где X – название переменной первой выборки,

Y – название переменной второй выборки.

Для проведения теста знаков Фишера о равенстве медиан для двух выборок в статистическом пакете STATA используется команда signtest X=Y

Для проведения теста Крускалла-Уоллеса о равенстве медиан нескольких выборок в статистическом пакете STATA используется команда

kwallis varname, by(groupvar),

где varname – имя переменной, объединяющей наблюдения из всех выборок,

groupvar – переменная, характеризующая принадлежность к каждой из рассматриваемых выборок..

Для проведения теста Колмогорова-Смирнова на равенство распределений в статистическом пакете R используется команда

ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"), exact = NULL),

где x и y – имена переменных, для которых сравниваются распределения,

alternative – вид альтернативной гипотезы ("two.sided", "less" или "greater"),

exact – ответ на вопрос, требуется ли вычислять p-value (не используется в случае если есть повторяющиеся наблюдения).

Для проведения теста Уилкоксона – Манна - Уитни гипотезы о равенстве средних для двух выборок в статистическом пакете R используется команда

wilcox.exact(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, exact = NULL, conf.int = FALSE, conf.level = 0.95, ...)

где x и y – имена переменных, для которых сравниваются медианы,

mu - значение средних, для которой проверяется тест, по умолчанию равно 0,

alternative – вид альтернативной гипотезы ("two.sided", "less" или "greater"),

conf.level – уровень доверия, например, 0.95.

Для проведения теста знаков Фишера о равенстве медиан для двух выборок в статистическом пакете R используется команда

SIGN.test(x, y = NULL, md = 0, alternative = "two.sided", conf.level = …),

где x и y – имена переменных, для которых сравниваются медианы,

md - значение медианы, для которой проверяется тест, по умолчанию равно 0,

alternative – вид альтернативной гипотезы ("two.sided", "less" или "greater"),

conf.level – уровень доверия, например, 0.95.

Для проведения теста Крускалла-Уоллеса о равенстве медиан нескольких выборок в статистическом пакете R используются команды

kruskal.test(x, ...)

kruskal.test(formula, data, subset, na.action, ...),

где x - вектор значений или список векторов значений,

formula - данные можно передавать в виде пары: “значение” ~ “группирующая переменная”,

... - прочие переменные, в основном настройки вывода и фильтров на данные.

Упражнения на применение непараметрических тестов.

Упражнение 1. В файле miles содержатся данные о количестве миль в расчете на единицу израсходованного топлива, проходимых 12 машинами одного класса при одинаковом режиме эксплуатации при использовании «обычного» топлива (miles1) и топлива со специальными добавками (miles2). Есть ли статистически значимая разница (уровень значимости выберите самостоятельно)? Для ответа на поставленный вопрос примените а) критерий Колмогорова-Смирнова и б) тест знаков Фишера.

Упражнение 2. В файле Change_Job содержатся, в том числе, данные о месячном доходе мужчин и женщин

Переменные:

changejob – ответ на вопрос: «Хотите ли Вы сменить работу?» (0 – нет, 1 – да),

age - возраст респондента в 2006 г.,

sex – пол респондента (1 – м, 2 – ж),

income – сколько денег получил респондент за последние 30 дней,

boss – ответ на вопрос: «Есть ли у Вас подчиненные?» (1 – да, 2 – нет),

С помощью критерия критерий суммы рангов Уилкоксона-Манна-Уитни сравните средний доход мужчин и женщин, желательно одного (или близкого) возраста. Соответствующие переменные создайте самостоятельно.

Упражнение 3. В файле Youth_unemployment содержатся данные об уровне безработицы в группе 20-29 летних в семи федеральных округах России в 1997-2008 гг. Выберите для исследования один год. Применяя критерий Крускалла-Уоллеса, проверьте, можно ли считать средний уровень безработицы в Федеральных округах одинаковым?

Упражнение 4. Продемонстрируйте применение перечисленных выше непараметрических критериев при анализе самостоятельно выбранных Вами данных (можно, например, использовать данные RLMS с сайта НИУ-ВШЭ или Росстата (gks.ru)).

Литература

Hollander M., Wolfe D., “Nonparametric statistical methods”, New York, John Wiley & Sons, 1999. (HW)

1) Lehmann, Erich L., “Nonparametrics: Statistical Methods Based on Ranks”, Springer, 2006.
2) Айвазян С.А., Мхитарян В.С., «Прикладная статистика и основы эконометрики», М.: Юнити, 1998.