«Методы оценки функции распределения доходов населения»
Студентка 4-го курса Пенухина Елена

Цель доклада: обзор работ по методам оценки функции распределения доходов населения.

План выступления:

1. Краткая история вопроса
2. Описание нескольких работ зарубежных авторов по выбранной теме
3. Краткое описание методологии Росстата
4. Подробный разбор модели С.А. Айвазяна
5. Анализ реальных данных для России

Краткая история вопроса

1897 г. – Работа В.Парето «Le cours d'economie politique»
Функция плотности вероятности f(*x*) дохода *x* имеет вид:

$$f(x) = \frac{\alpha}{c}\left(\frac{c}{x}\right)^{\alpha+1}, x \geq c \quad (1),$$

где c – константа, а α – индекс Парето.
Чем меньшее значение принимает α, тем более неравномерно распределены доходы.
1922 г. – Опубликована работа Джини, в которой он также показал, что распределение доходов соответствует степенному закону, но с неуниверсальными показателями степени.
1931 г. – Жибра в работе «Les inegalites economique» выдвинул гипотезу о том, что доходы населения подчиняются логарифмически нормальному распределению.
Функция плотности вероятности:

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{\frac{-(\ln x - \ln a)^2}{2\sigma^2}} \quad (2),$$ где a – математическое ожидание, а σ^2 - дисперсия.

$$\beta \equiv \frac{1}{\sqrt{2\sigma^2}} \quad (3)$$ называют индексом Жибра. Небольшое значение индекса β соответствует неравномерному распределению доходов.
1960 г. – Мандельброт предложил использовать «слабый закон Парето», показав, что закон Парето применим только асимптотически на концах распределений.

Обзор зарубежных работ по методам оценки функции распределения

A.Banerjee, V.M.Yakovenko, T. Di Matteo «A study on the personal income distribution in Australia», Physica A 370, 2006.
Исследуемая страна: Австралия
Источник данных: Australian Bureau of Statistics
Период: 1989 – 2000 гг.
Выборка: примерно 14000 репрезентативных индивидов

$$P(x) = \begin{cases} \frac{1}{T}\exp(-x/T) & \text{экспоненциальное распределение,} \\ \frac{1}{xs\sqrt{2\pi}}\exp\left[\frac{-\log^2(x/m)}{2s^2}\right] & \text{логнормальное распределение,} \\ \frac{(\beta)^{-(1+\alpha)}}{\Gamma(1+\alpha,0)} x^{\alpha}\exp(-x/\beta) & \text{гамма распределение.} \end{cases} \quad (4)$$

Интегральная функция распределения: $C(x) = \int_x^\infty P(x')dx'$ (5)

При оценивании рассматривались только значения C(x) от 100 до 1%.

Год	T (тыс.$)	m (тыс.$)	s	β (тыс.$)	α	σ (%)			Мода ($)
						Экспон.	Логнорм.	Гамма	
1989-1990	17.8	15.1	0.74	13.4	0.39	13	11	6.8	6196
1993-1994	18.5	18.8	0.63	13.1	0.59	18	9.6	5.7	7020
1994-1995	19.6	17.7	0.71	14.9	0.40	15	9.4	5.5	7280
1995-1996	20.5	18.2	0.72	15.7	0.39	14	8.6	6.5	7280
1996-1997	21.2	18.9	0.72	16.5	0.37	14	8.4	7.7	7540
1998-1999	23.7	19.0	0.79	19.6	0.25	10	11	7.1	7800
1999-2000	24.2	19.6	0.78	19.3	0.30	11	11	7.2	7800

Табл.1. Параметры распределений (3)

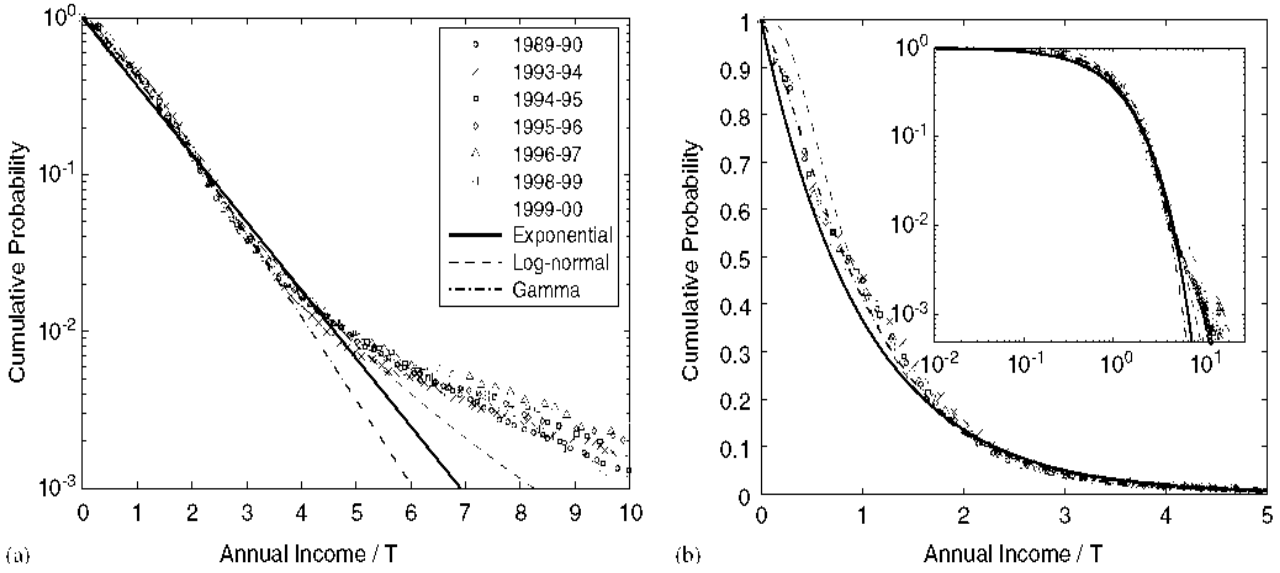

График 1. Интегральная функция распределения дохода.

Выводы:

1. *Экспоненциальное распределение* при аппроксимации доходов является предпочтительным, так как оно достаточно точно приближает реальные данные для 98% населения и имеет всего один параметр. Логнормальное и гамма распределения незначительно улучшают приближение, но при этом имеют большее число параметров, что усложняет анализ.
2. Отклонения реальных данных от теоретического экспоненциального распределения рассматриваются авторами работы как результат проводимой государством политики перераспределения доходов.
3. Среднеквадратическое отклонение эмпирической плотности вероятности от теоретической предлагается рассматривать как меру эффективности проводимой политики.

F.Clementi, M.Gallegati «Power law in the Italian personal income distribution», 2005

Исследуемая страна: Италия

Источник данных: Survey on household income and wealth (Bank of Italy), National Institute of Statistics

Период: 1977 – 2002 гг.

Выборка: примерно 10000 индивидов

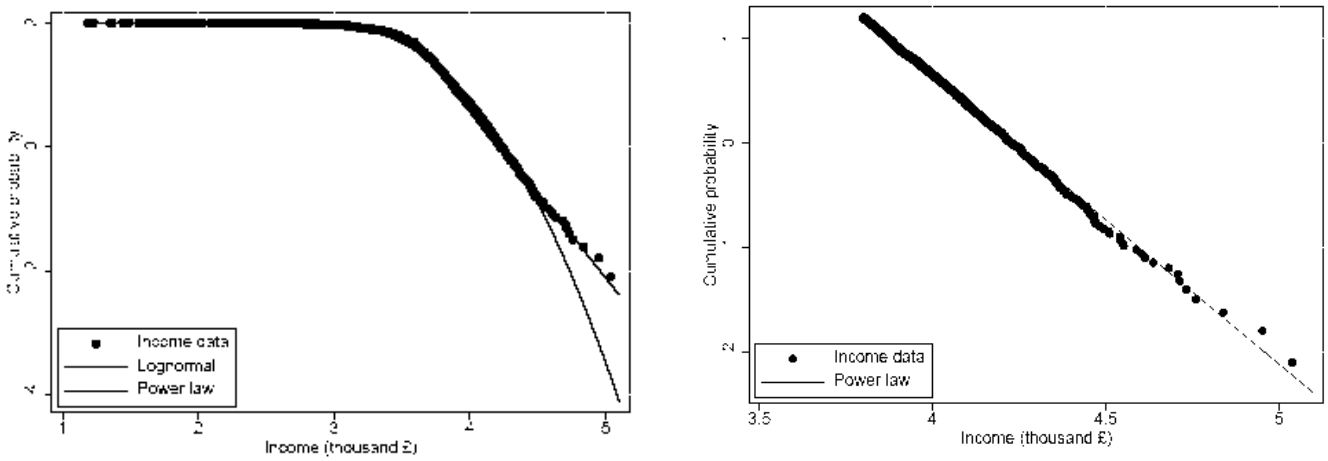

График 2-3. Интегральная функция распределения доходов в Италии в 1998 г.

В качестве законов распределения, наилучшим образом аппроксимирующих реальные данные, были выбраны логарифмически нормальное распределение для участка с низкими и средними доходами и Парето для хвоста распределения.

Все доходы приведены с помощью ИПЦ к сопоставимым ценам 1976 г.

Для оценивания параметров логнормального распределения был использован *метод максимального правдоподобия*. Параметры Парето распределения были найдены *методом наименьших квадратов*.

Год	ln(a)	σ	α	c	R^2
1977	3.31 (0.005)	0.34 (0.004)	3.00 (0.008)	10876	0.9921
1983	3.38 (0.005)	0.30 (0.005)	3.11 (0.006)	11147	0.9945
1989	3.53 (0.003)	0.26 (0.003)	2.91 (0.002)	15788	0.9995
1995	3.46 (0.004)	0.32 (0.003)	2.72 (0.002)	16587	0.9996
1998	3.48 (0.004)	0.34 (0.006)	2.76 (0.002)	17141	0.9993
2002	3.52 (0.004)	0.31 (0.005)	2.71 (0.002)	17664	0.9997

Табл.2. Параметры распределений

Выводы: В Италии распределение доходов населения удовлетворяет логнормальному закону на области низких и средних доходов и закону Парето на области высоких доходов (примерно 1% населения).

Методология Росстата

Единицей обследования является *домашнее хозяйство*.

Выборка для проведения обследования бюджетов домашних хозяйств формируется на принципах представительности категории *«все население»* в пределах каждого региона РФ.

Статистическое взвешивание результатов.

1) расчет «базовых» весов, позволяющих привести данные выборочного обследования к генеральной совокупности, исходя из общих принципов отбора;

2) корректировка «базовых» весов выборки на смещение, которое возникает из-за невозможности получения полной информации по всем домашним хозяйствам, входящим в выборку.

Для расчета «базовых» весов на первом этапе взвешивания Росстатом используется следующая информация:

- данные микропереписи населения 1994 года о распределении членов домашних хозяйств по семьям различного состава отдельно по городскому и сельскому населению в региональном разрезе;
- данные о численности населения по состоянию на 1 января текущего года для городского и сельского населения в региональном разрезе.

Корректировка на втором этапе взвешивания проводится на основе использования региональных данных о среднедушевых денежных доходах генеральной совокупности и гипотезы о соответствии распределения доходов в генеральной совокупности *закону логнормального распределения*.

Модель С.А. Айвазяна

С.А. Айвазян, С.О. Колеников, «Уровень бедности и дифференциация по расходам населения России», итоговый отчет, Российская программа экономических исследований, 2000
Источник данных: 5-8 раунды RLMS и статистика Бюджетных обследований домашних хозяйств (БОДХ)

Основные источники искажения статистических данных:
1. Отказ домашнего хозяйства от участия в обследовании
2. Намеренное искажение данных при опросе, с целью сокрытия неофициальных источников дохода
3. Невозможность обследования домашних хозяйств, чей доход превышает некоторый критический уровень

Предмодельные гипотезы:
H1*. Распределение среднедушевых доходов домашних хозяйств может быть адекватно описано смесью логарифмически нормальных законов.*

Постулат 1.

Распределение населения по среднедушевым доходам ξ внутри однородной страты (по источникам доходов, территориальным, социальным, профессиональным и демографическим признакам) подчинено логнормальному закону распределения с параметрами

$a_j = E(\ln \xi_j)$

$\sigma_j^2 = D(\ln \xi_j)$, где ξ_j – среднедушевые расходы случайно извлеченного домашнего хозяйства j-ой однородной страты.

Постулат 2.

Если считать, что общество состоит из непрерывного по средней величине логарифмов расходов спектра страт, то при некотором естественном виде смешивающей функции q(a) распределение всего населения также будет подчиняться логарифмически нормальному закону.

Постулат 3.

При нарушении непрерывности спектра страт или при нарушении условия монотонного убывания смешивающей функции q(a) общее логнормальное распределение трансформируется в смесь логнормальных законов.

H2*. Вероятность уклонения домашнего хозяйства от участия в бюджетном обследовании является функцией от некоторых его социально-экономических и территориальных характеристик.*
H3. *Коэффициент вариации среднедушевых расходов домашних хозяйств является постоянным, т.е. не зависит от номера социально-экономической страты, для которой он вычисляется.*
H4. *Распределение расходов в статистически ненаблюдаемой «супер-богатой» страте может быть описано трехпараметрическим логнормальным законом, оцененным по наблюдениям, относящимся к статистически обследованным стратам населения.*

Описание модели.

Плотность распределения случайной величины ξ будет описываться моделью смеси логнормальных законов распределения следующего вида:

$$f(x|\Theta)=\sum_{j=1}^{k} q_j \frac{1}{\sqrt{2\pi}\sigma_j x} e^{-\frac{(\ln x - a_j)^2}{2\sigma_j^2}} + q_{k+1}\frac{1}{\sqrt{2\pi}\sigma_{k+1}(x-x_0)} e^{-\frac{(\ln(x-x_0)-a_{k+1})^2}{2\sigma_{k+1}^2}}, \quad (6)$$

где ξ (тыс. руб.) – среднедушевой расход случайно извлеченного из генеральной совокупности индивида;

$\Theta = (k; q_1,...,q_{k+1}; a_1,...,a_{k+1}; \sigma_1^2,...,\sigma_{k+1}^2)$ - параметры модели, имеющие следующее содержание:

k+1 – число однородных социально-экономических страт в обществе;

qj (j=1,2,…,k+1) – априорная вероятность появления наблюдений, представляющих j-ю однородную страту;

$a_j = E(\ln \xi_j)$ (j=1,2,…,k+1) – это теоретические средние значения логарифмов среднедушевых расходов внутри j-й страты;

$\sigma_j^2 = D(\ln \xi_j)$ (j=1,2,…,k+1) – дисперсии логарифмов среднедушевых расходов внутри однородной социально-экономической страты;

x0 – некоторое пороговое значение среднедушевых расходов, отделяющее статистически доступный диапазон расходов (x<x0) от статистически недоступного (x>x0).

Смесь законов распределения представляет собой сумму законов распределения каждой из однородных социально-экономических страт, где каждый из законов взвешен с помощью смешивающей функции.

Вероятность уклонения домашнего хозяйства от обследования

$$p(x) = P\{\eta_i = 0 | Z\} = \frac{1}{1+e^{\beta^T Z}}, \quad (7)$$

где η_i принимает два возможных значения:

0, если i-е домашнее хозяйство отказалось от участия в обследовании;

1, в противоположном случае;

Z=(1,z), где z=ln ξ – логарифм совокупных душевых расходов домашних хозяйств;

$\beta=(\beta 0, \beta 1)^T$ – вектор столбец искомых параметров модели.

Калибровка исходных данных

В исходной выборке данные калибруются следующим образом:

Наблюденные значения x	x_1	x_2	…	x_n
Веса наблюденных значений	$\frac{1}{n}$	$\frac{1}{n}$	**…**	$\frac{1}{n}$

Табл. 3. Исходная калибровка наблюдений

При условии, что никто не станет уклоняться от обследования и не станет искажать свои расходы число наблюдений ν(x*), попавших в малую Δ окрестность точки x* будет равно:

$$\mathbf{\nu(x^*) \approx n f(x^*) \Delta}, \quad (8)$$

где f(x) – функция плотности распределения населения по среднедушевым расходам,

n – число статистически обследованных индивидов,

x* - заданное значение среднедушевых расходов.

Реальное (т.е. наблюденное в выборке объема n) число наблюдений в Δ окрестности точки x*:

$$\tilde{\nu}(x^*) \approx nf(x^*)(1-p(x^*))\Delta . \quad (9)$$

$$\nu(x^*) = \tilde{\nu}(x^*) \cdot \frac{1}{1-p(x^*)}. \quad (10)$$

При достаточно малых значениях Δ:

$$\tilde{\nu}(x_i) = 1 \quad (11)$$

Теоретическое число наблюдений в Δ окрестности будет

$$\nu(x_i) = \frac{1}{1-p(x_i)}. \quad (12)$$

Наблюденные значения x	x_1	x_2	…	x_n
Веса наблюденных значений	ω_1	ω_2	**…**	ω_n

Табл.4. Калибровка, предложенная в модели

где вес ωi определяется как

$$\omega_i = \frac{\frac{1}{1-p(x_i)}}{\sum_{j=1}^{n}\left(\frac{1}{1-p(x_i)}\right)}. \quad (13)$$

Наблюдению придается тем больший вес, чем больше для него вероятность уклониться от обследования. Сумма всех весов равняется единице.

Оценка параметров статистически наблюдаемых компонентов смеси логнормальных распределений

Оценка параметров $k, \tilde{q}_1,...,\tilde{q}_k, a_1,...,a_k, \sigma_1^2,...,\sigma_k^2$ статистически наблюдаемых компонентов в смеси *логарифмически нормальных* законов распределения:

$$\tilde{f}(x) = \sum_{j=1}^{k} \tilde{q}_j \cdot \frac{1}{\sqrt{2\pi}\sigma_j x} e^{-\frac{(\ln x - a_j)^2}{2\sigma_j^2}} . \quad (14)$$

Задача была сведена к оценке тех же параметров в смеси *нормальных* распределений вида

$$\tilde{\varphi}(x) = \sum_{j=1}^{k} \tilde{q}_j \cdot \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(z-a_j)^2}{2\sigma_j^2}} , \quad (15)$$

где z=ln(x)

Оценивание проводится с помощью метода максимального правдоподобия.

Необходимо найти такие значения параметров $\hat{\Theta}(k) = (\hat{\tilde{q}}_1,...,\hat{\tilde{q}}_k;\hat{a}_1,...,\hat{a}_k;\hat{\sigma}_1^2,...,\hat{\sigma}_k^2)$, при которых логарифмическая функция правдоподобия

$$l_k(\Theta(k)) = \sum_{i=1}^{n} \omega_i \left[\ln \sum_{j=1}^{k} \tilde{q}_j \varphi(z_i | a_j; \sigma_j^2) \right] \quad (16)$$

достигает максимального значения.

При осуществлении оценивания параметров Θ считается, что число страт k задано.

Итерационный механизм определения числа k.

Последовательная проверка гипотез

H₀: k=j

H₁: k=j+1, j=1,2,…,

Критическая статистика:

$$\gamma(j) = -2\ln \frac{l_j(\hat{\Theta}(j))}{l_{j+1}(\hat{\Theta}(j+1))}. \quad (17)$$

Оценка параметров статистически ненаблюдаемых компонентов смеси и всего распределения.

Пусть удельный вес ненаблюдаемого $(\hat{k}+1)$-го компонента смеси равен $q_{\hat{k}+1}$, а среднее логарифмов среднедушевых расходов статистически ненаблюдаемой страты равно $a_{\hat{k}+1}$.

Тогда общее среднее μ по всей генеральной совокупности может быть вычислено по формуле математического ожидания, исходя из известного вида общего для всего населения закона распределения:

$$\mu = \int_0^{\infty} x \left(\sum_{j=1}^{\hat{k}} \hat{q}_j \frac{1}{\sqrt{2\pi}\hat{\sigma}_j x} e^{-\frac{(\ln(x)-\hat{a}_j)^2}{2\hat{\sigma}_j^2}} + q_{\hat{k}+1} \frac{1}{\sqrt{2\pi}\sigma_{\hat{k}+1}(x-x_0)} e^{-\frac{(\ln(x-x_0)-a_{\hat{k}+1})^2}{2\sigma_{\hat{k}+1}^2}} \right) dx, \quad (18)$$

где $\hat{q}_j = \hat{\tilde{q}}_j(1-q_{\hat{k}+1})$, j=1,2,…, $\hat{k}$. (19)

По свойствам логнормального распределения:

$$\mu = \sum_{j=1}^{\hat{k}} \hat{q}_j e^{\frac{\hat{\sigma}_j^2}{2}+\hat{a}_j} + q_{\hat{k}+1}\left(x_0 + e^{\frac{\sigma_{\hat{k}+1}^2}{2}+a_{\hat{k}+1}} \right). \quad (20)$$

x0 – параметр сдвига.

Величина μ зависит от четырех неизвестных параметров: $q_{\hat{k}+1}$, $a_{\hat{k}+1}$, x0, и $\sigma_{\hat{k}+1}^2$.

По заданным в начале предпосылкам величина x0 может быть определена как максимальное наблюденное значение расходов в выборке:

$$x_0 = \max_{1\le i\le n}\{x_i\} \quad (21)$$

Согласно третьей гипотезе, можно вычислить общую оценку для величины σ^2, представив ее в виде суммы дисперсий каждой из страт, взвешенных с помощью смешивающей функции.

$$\hat{\sigma}^2 = \sum_{j=1}^{\hat{k}} \hat{q}_j \hat{\sigma}_j^2 . \qquad (22)$$

По третьей гипотезе предполагается, что величина $\sigma_{\hat{k}+1}^2$ равна $\hat{\sigma}^2$.

Для оценки оставшихся двух параметров в системе координат ($q_{\hat{k}+1}, a_{\hat{k}+1}$) вычисляются линии уровня исходя из того, что

$$\mu(q_{\hat{k}+1}, a_{\hat{k}+1}) = \mu^{\text{макро}} \qquad (23)$$

$\mu^{\text{макро}}$ – среднее значение среднедушевых расходов, полученное из «Балансов доходов и расходов населения», публикуемых Росстатом.

Конкретный выбор точки ($\hat{q}_{\hat{k}+1}, \hat{a}_{\hat{k}+1}$) на линии уровня определяется экспертным путем.

Выводы:
1. модель Айвазяна позволяет учесть специфику российской функции распределения, а именно отсутствие среднего класса, а как следствие – наличие второго горба в функции распределения доходов;
2. модель вносит существенный вклад в развитие косвенных методов оценки функции распределения доходов населения;
3. модель не совершенна.

Анализ реальных данных для России.

Источник данных: Федеральная служба государственной статистики
Период: 2000 – 2005 гг.

- Плотность логнормального распределения:

$$p_\eta(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\ln x - \ln a)^2}{2\sigma^2}} \qquad (24)$$

- Функция распределения:

$$F_\eta(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\ln x} e^{\frac{-(t-\ln a)^2}{2\sigma^2}} dt \qquad (25)$$

- Медиана:

$$Me(\eta) = a \qquad (26)$$

- Среднеквадратическое отклонение

$$\sigma_{\ln x} = \sqrt{\frac{\sum (\ln x_i - \overline{\ln x_i})^2 w_i}{\sum w_i}} \qquad (27)$$

С помощью формул (26) и (27) были вычислены значения медианы и среднеквадратического отклонения.

	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Среднеквадратическое отклонение логнормального распределения	0,654	0,702	0,703	0,728	0,721	0,682
Медиана	3401	3790	4149	4865	5355	6079

Табл.5. Результаты вычислений

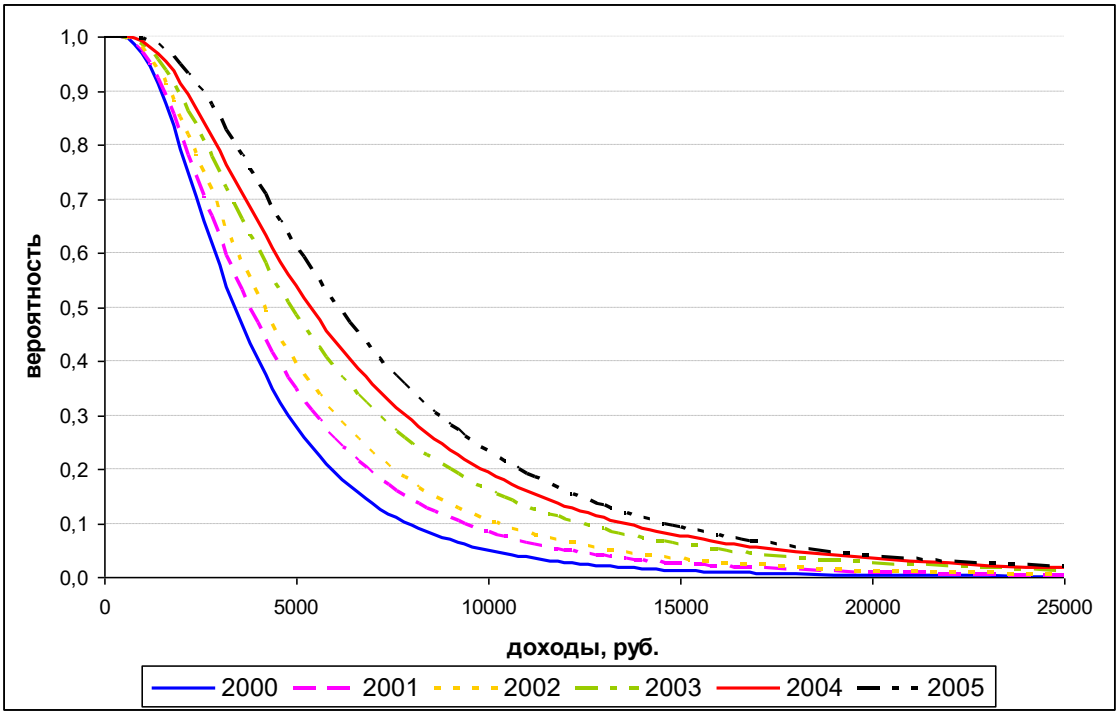

График 4. Интегральная функция распределения доходов населения

На графике прослеживаются две противоположные тенденции:
1. Общий рост реальных доходов населения
2. Рост дифференциации

	2000	2001	2002	2003	2004	2005
Коэффициент фондов	13,9	13,9	14,0	14,5	15,2	14,8
Коэффициент Джини	0,395	0,397	0,397	0,403	0,409	0,405
Децильный коэффициент	5,926	6,856	6,894	7,198	6,963	6,235

Табл.5. Коэффициенты дифференциации[1]

Список литературы.

1. A.Banerjee, V.M.Yakovenko, T. Di Matteo «A study on the personal income distribution in Australia», Physica A 370, 2006.
2. F.Clementi, M.Gallegati «Power law in the Italian personal income distribution», 2005
3. С.А. Айвазян, С.О. Колеников, «Уровень бедности и дифференциация по расходам населения России», итоговый отчет, Российская программа экономических исследований, 2000
4. «Методологические положения по статистике», Федеральная служба государственной статистики
5. Статистический сборник «Россия в цифрах. 2006»: Стат.сб./Росстат.-М., 2006

[1] Коэффициент фондов и коэффициент Джини взяты с официального сайта Росстата: www.gks.ru/free_doc/2006/b06_13/06-01.htm