Trends in the Relative Distribution of Wages by Gender and Cohorts in Brazil (1981-2005)

Ana Maria Hermeto Camilo de Oliveira
Affiliation: CEDEPLAR/UFMG
Address: Av. Antônio Carlos, 6627 – FACE/UFMG – Belo Horizonte, MG, Brazil – 31270-901
Phone Number: 55-31-34097154
E-mail: ahermeto@cedeplar.ufmg.br

Raquel Rangel de Meireles Guimarães
Affiliation: CEDEPLAR/UFMG
Address: Av. Antônio Carlos, 6627 – FACE/UFMG – Belo Horizonte, MG, Brazil – 31270-901
Phone Number: 55-31-34097154
E-mail: rguima@cedeplar.ufmg.br

Extended Abstract

Most of previous works regarding gender wage differentials are based on parametric methodologies for estimation of wage differentials by gender, using traditional mincerian regressions or quantile regression estimates, with classical or bayesian statistics. The aim of our paper is to analyse and decompose changes in earnings relative distribution between men and women in different cohorts, using the relative distribution framework. This methodology was proposed by Handcock and Morris (1999), considering non-parametrical tools which allow an exploratory analysis that is independent of parametric assumptions on the mathematical form of the response-variable probabilities. We use density estimates of the kernel probability for each sex and cohort and decompositions of the relative distribution to get substantive evidences for gender differentials and relative mobility in Brazil, from 1981 to 2005.

We use microdata from the Brazilian Household Sample Survey (*Pesquisa Nacional por Amostra de Domicílios* - PNAD), for 1981, 1992 and 2005. This survey is yearly conducted by the Brazilian Census Bureau (*Instituto Brasileiro de Geografia e Estatística – IBGE*) and presents a comprehensive data source, in particular about the labor market and earnings, statistically representative of Brazil. To analyze the wage differentials between male and female workers, our sample was restricted to those working at the survey reference week and earned a positive wage in the period. We constructed a pseudo-panel of these repeated cross-sections, that allows us to follow cohorts over time.

Non-parametric approaches brought into consideration the fact that it was not strictly necessary to adopt assumptions about the mathematical form of the probabilities distribution of a variable. Most of parametric models (classical regressions and their decompositions) are sensitive to violations of their hypothesis, turning into misleading answers to studies questions (DINARDO e TOBIAS, 2001). Moreover, the non-parametric methodology allows the analysis of data as they are, without any prior distribution assumption.

Therefore, the relative distribution method, proposed by Handcock and Morris (1999), constitutes a valuable instrument for the substantive analysis of the wage inequality and provides a consistent framework for data analysis. Intuitively, the relative distribution is a transformation of data from two distributions (reference and comparison - in our paper, men and women) into one

distribution that contains all information for comparison between them. In addition, the relative distribution combines the exploratory potential of data to a statistical tool of estimation and inference that is insensitive to recurring assumptions of parametric methodologies.

Despite the development of data analysis based on the relative distribution and other non-parametric methods, there are few studies in Brazil that examine the evolution of gender wage inequality using these tools, particularly following cohorts over time. Besides, although inequality studies should focus on the distribution as a whole, most of empirical studies are stuck to methodologies based on mean values, which are not always representative of the population. The relative distribution method fills this lack and provides a framework of summary measures and figures that allows a substantive examination of data.

The intuition of the relative distribution is the construction of counterfactual settings in which two populations are compared, based on the probabilities distributions. In our paper, we investigate how female workers in Brazil would be located at the male wage distribution, or how they would be located in 2005 if their wage distribution remains as in 1997, for example.

The relative distribution has the following properties: (*i)* it is not affected by scale (invariant to any monotonic transformation of the original variable; e.g. wages versus log-wages); (*ii*) the basic unit of analysis is the population and not the individual; (*iii*) it measures individuals proportions and ranks, and not the earnings values, as the traditional methods. These characteristics distinguish the relative distribution for inequality questions.

Formally, for our empirical application, consider Y_0 as a random variable for the hourly-wage logarithm of a reference population (men). The probability density funcion (*pdf*) of Y_0 is given by $f_0(y)$ and its cumulative distribution (*cdf*) is $F_0(y)$. Consider also the same measures for a comparison population Y (private sector workers), with a *pdf* of Y, $f(y)$ and its distribution, $F(y)$. The relative distribution of Y to Y_0 (*R function*) is defined by the random variable distribution:

$$R = F_0(Y) \qquad (1)$$

Where:

R: indicates the relative distribution of log hourly-wages,

F_0: is the cumulative distribution function for the log hourly-wages of male workers, and

Y: is the log hourly-wages of female workers.

The *R* relative distribution is obtained from *Y*, changed by the cumulative distribution function for Y_0, F_0. A property of *R* is that it is continuous at the [0,1] range. We define the outcome of *R*, *r*, by relative data.

An important measure related to the cumulative distribution functions (*cdf*) is their inverse function that derives the quantile function. For the log hourly-wages of a population, this function is described as:

$$Q(p) = F^{-1}(p) = \inf_x \{x | F(x) \geq p\} \quad (2)$$

Where:

Q(.): indicates the quantile function of the log hourly-wages,

$F^{-1}(.)$*:* is the inverse of the cumulative distributive function (*cdf)* of hourly-wages,

x: is the observed hourly-wages, and

p: indicates the quantile of interest.

The quantile function value, *Q(p),* defines the *p* quantile value of the hourly-wage distribution. An special case is the median *(p = 0,5)*, for which the *Q(0,5)* value defines the hourly-wage that splits 50% of the population whose wage is below e 50% above.

Given that, by definition, the relative distribution is a monotonic transformation of the variable of interest (log hourly-wages for relative data) and, if we know this variable distribution, we are able to show the equivalence between the original distribution and its transformation:

$$F_Y(y) = P(X \leq h^{-1}(x)) = F(h^{-1}(y)) \quad (3)$$

Where:

Y: is the variable of interest, and

h(x): indicates the monotonic transformation of this variable.

Hence, the equation (3) proves that the cumulative distribution function of *Y* is equivalent to the same function for a monotonic transformation of *Y*. Using this property, we can reach the cumulative distribution function (*cdf)* of the random variable *R*:

$$G(r) = F(F_0^{-1}(r)) = F(Q_0(r)), \quad 0 \leq r \leq 1 \quad (4)$$

The first derivate of G(r) indicates the relative density function (*pdf*):

$$g(r) = \frac{f(Q_0(r))}{f_0(Q_0(r))}, \quad 0 \leq r \leq 1 \quad (5)$$

So, the relative density *g(r)* can be interpreted as a density ratio: the ratio between the fraction of female workers and the fraction of male workers at a given level of the outcome *Y* ($Q_0(r)$). Intuitively, the relative density points how the women would be located at the wage distribution of men. If wages over the distribution for the both sectors were equal, the *g(r)* function would assume the value 1 and, for example, the 10% of male workers that are located in the lower tail of the distribution would be equivalent to the 10% of lowest wages of the female distribution (if *g(0,10) = 1*).

To estimate the hourly-wage probabilities densities for the men and women, $f_0(y)$ e $f(y)$, we used the *kernel* non-parametric data smoothing method. This method allowed us to not adopt an unnecessary or wrong assumption about the wage distribution mathematical form.

The estimation of *kernel* densities require two components: the bandwidth and the *kernel* function. The bandwidth indicates the distance from a point x_0 of the wage distribution to where the density will be smoothed and the *kernel* is the function that estimates local means. According to Dinardo and Tobias (2001), the *kernel* choice does not significantly affects the form of the estimated probabilities curves, but the bandwidth choice has significant implications on the results, since it involves a *trade-off* between bias and variance. We used the optimal criteria proposed by Silverman (1986), which adopts an Epanechnikov kernel ($k(x)$) for the estimation of the probability density ($\hat{f}_n(x)$), with a bandwidth (h^*) defined as:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}k\left(\frac{x - X_i}{h}\right) \qquad (6)$$

$$k(x) = \frac{3}{4}\left(1 - x^2\right)I(x) \qquad (7)$$

$$h^* = 1{,}3643\ \delta n^{-0{,}2}\sigma \qquad (8)$$

Where:

x: is the function value to be smoothed,

k(x): is the kernel function,

I(x): is the index function,

h^* : is the optimal bandwidth,

δ : is a constant for the Epanechnikov kernel (1,7188),

n: is the sample size, and

σ : is the sample standard-deviation.

In addition to the easy and flexible interpretation, the relative distribution also provides decomposition methods of shifts in the median (location) and in the structure (shape), allowing interesting analyses about the effect over time of changes on wages. All this is connected to the explanatory power of the figures, which facilitates the interpretations even more.

Following Handcock e Morris (1999), consider Y_{0L} as a random variable that describes an outcome (here, standardized wages) for the reference group, adjusted to have the same median of the comparison group, i.e. $Y_{0L} = Y_0 - \rho$, where ρ is the variable median for the comparison group. In this case, we say that Y_{0L} is an hypothetical group that has the comparison group median, but the structure of the reference group. The cumulative distribution function of Y_{0L} is $F_{0L}(y)$ *or* $F_0(y - \rho)$. In the same way, the probability density of Y_{0L} is given by $f_{0L}(y)$ *or* $f_0(y - \rho)$.

For the relative decomposition in location (median) and shape (structure) shifts, we constructed two relative distributions, derived from the distributions of three groups: the reference group Y, the comparison group Y_0, and the hypothetical group, Y_{0L}, which has the median of the comparison group, but the structure of the reference group.

Consider R the relative distribution of the reference group to the comparison group, that is, $F_0(Y)$. Also consider R_0^{0L} as the relative distribution of the hypothetical group to the comparison group, $F_0(Y_{0L})$ or $F_0(Y_0 + \rho)$. In this case, if the reference and the comparison groups have the same median, then R_0^{0L} has an uniform distribution.

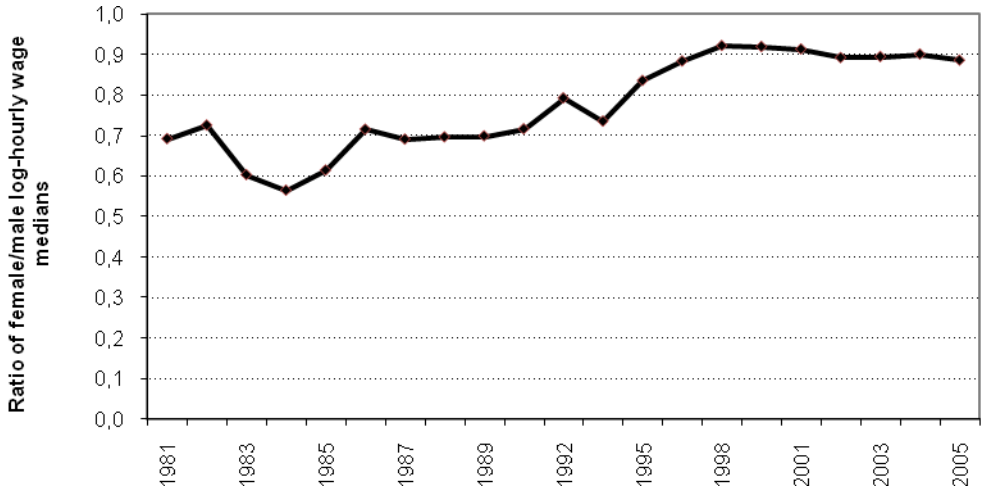
Considering R_{0L} as the relative distribution of the reference group to the hypothetical group, $F_{0L}(Y)$ or $F_0(Y - \rho)$, R_{0L} would have an uniform distribution when, net of median shifts, the two distributions (reference and comparison) have the same structure.

Therefore, the overall relative density $g(r)$ can be decomposed in two parts: a relative density that indicates location shifts $g_0^{0L}(r)$ and a relative density for differences in the two distributions shapes $g_{0L}(r)$:

$$g(r) = g_0^{0L}(r) \times g_{0L}(r) \qquad (9)$$

An application of this decomposition refers to the analysis of changes in the wages distribution between men and women in two periods.

Figure 1: Gender log-hourly wage gap. Brazil, 1981-2005



Source: Microdata extracted from Pesquisa Nacional por Amostra de

Figure 2: Relative distribution of the log-hourly wages by gender, Brazil, 1981-2005

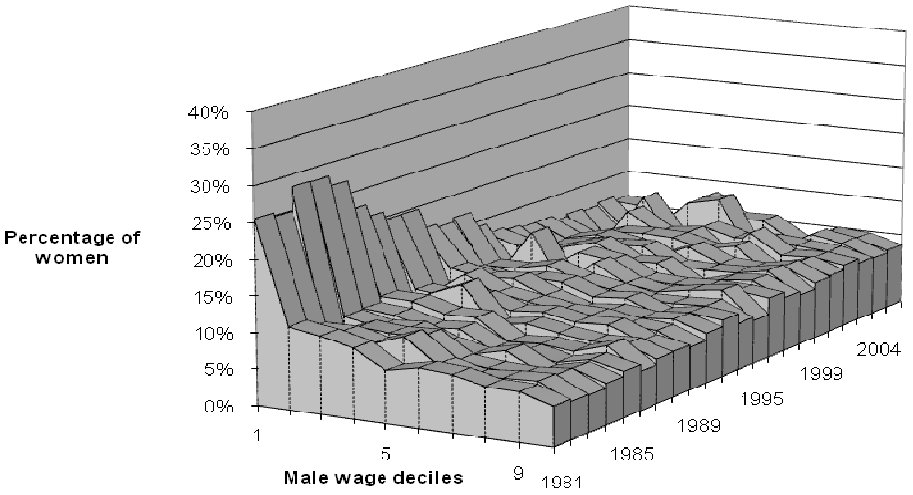
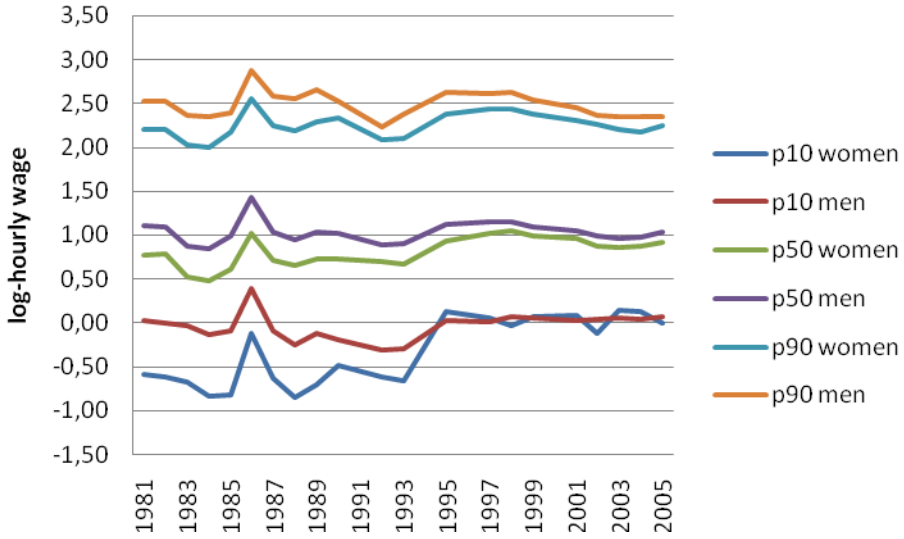
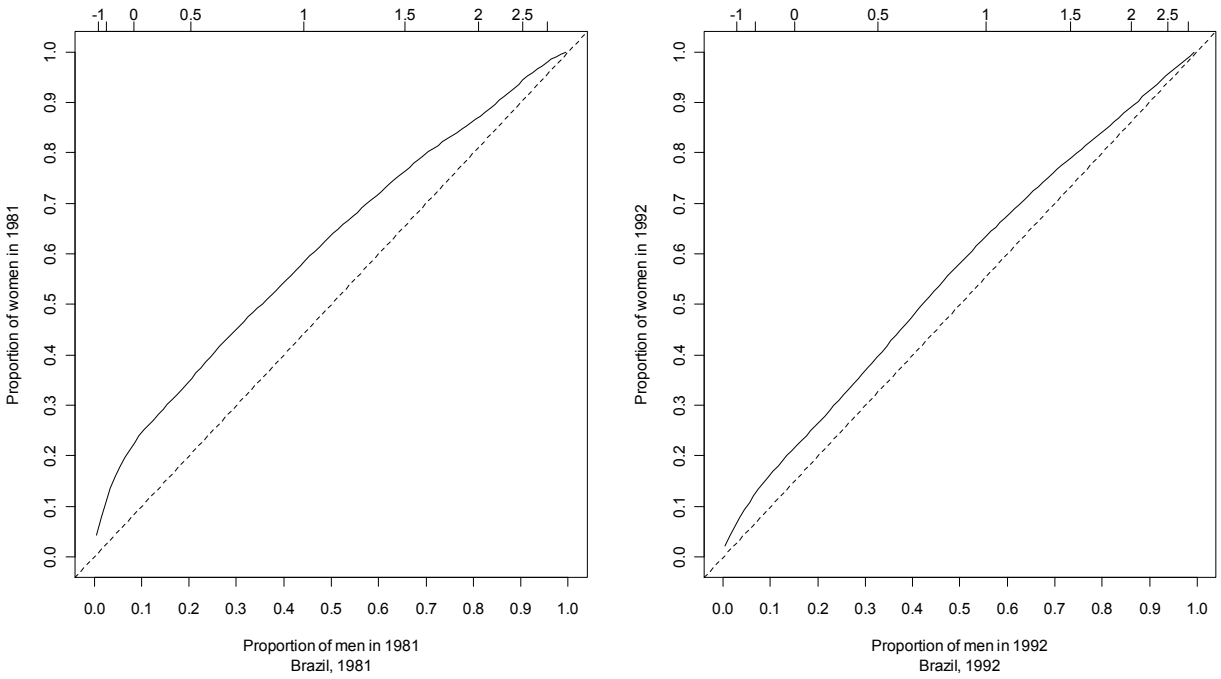


Figure 3: Selected deciles of log-hourly wage by gender, Brazil, 1981-2005
Source: Microdata extracted from Pesquisa Nacional por Amostra de



Source: Microdata extracted from Pesquisa Nacional por Amostra de

Figure 4: CDF of the Relative Distribution of wages by gender, Brazil, 1981 and 1992



Source: Microdata extracted from Pesquisa Nacional por Amostra de

Following Handcock and Aldrich (2002), if the two distributions (of men and women) are identical, then the cumulated distribution function (cdf) of the relative distribution line is a 45^{o} line and the probability density function (pdf) is that of a uniform distribution. Then we can infer that the gender distribution of wages in Brazil has becomed more equal since 1981.

Figure 5: CDF of the Relative Distribution of wages by gender, Brazil, 1992 and 2005

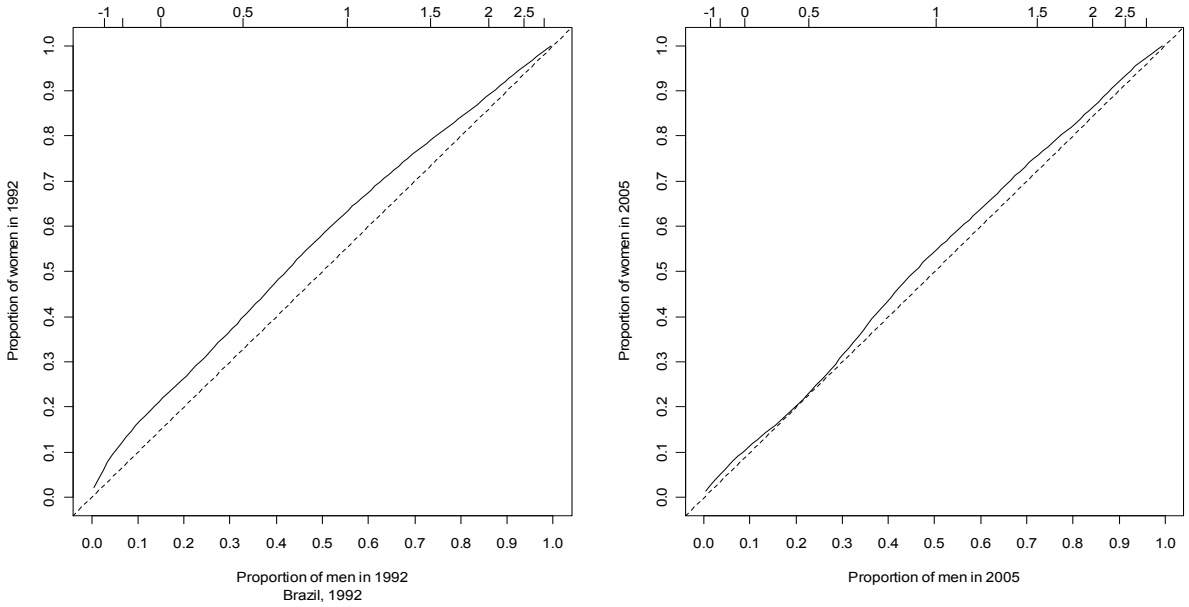
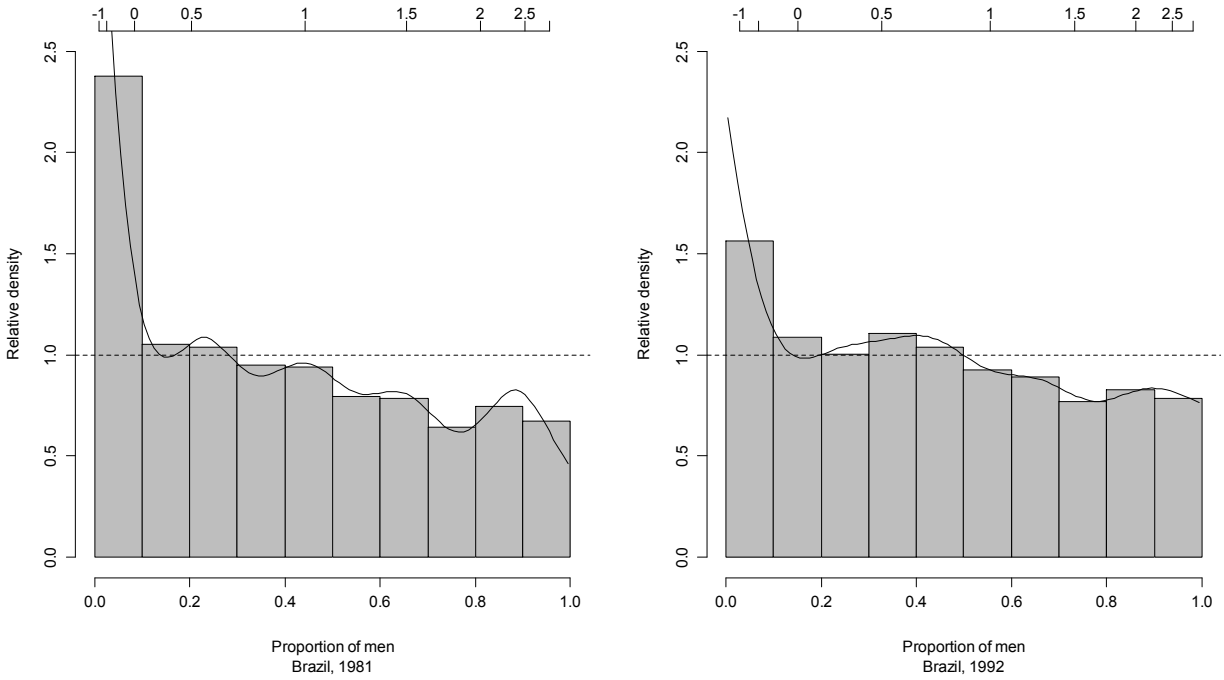
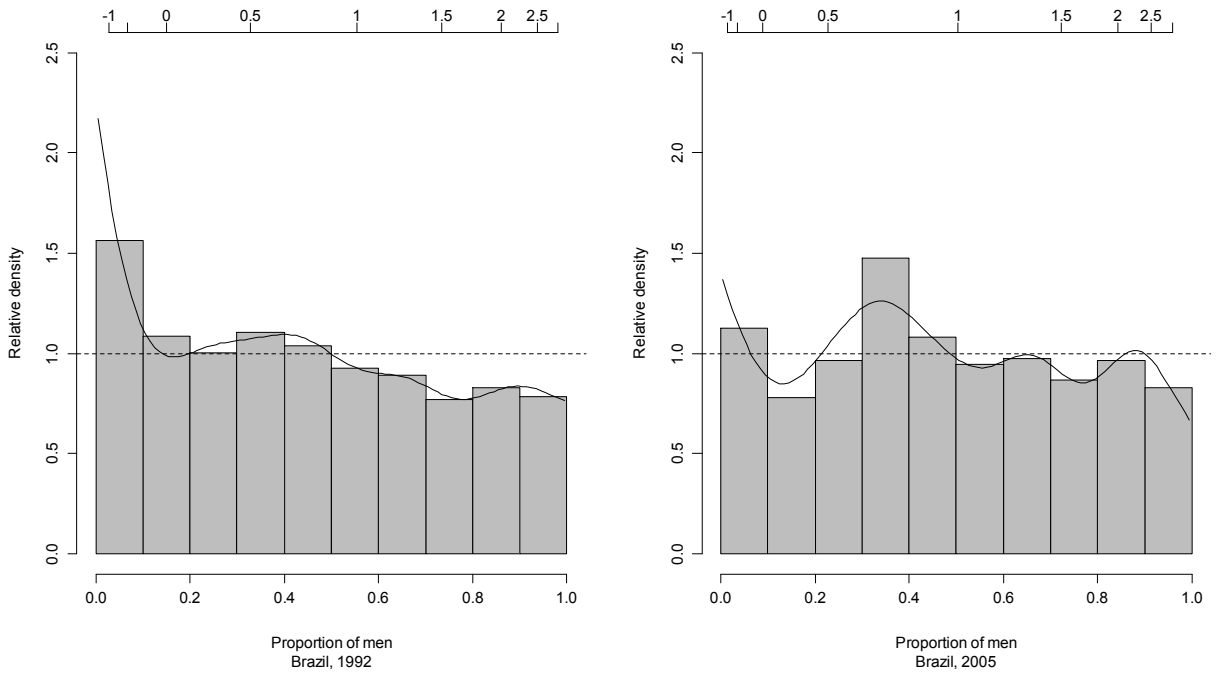


Figure 6: PDF of the Relative Distribution of wages by gender, Brazil, 1981 and 1992



Source: Microdata extracted from Pesquisa Nacional por Amostra de

Figure 7: PDF of the Relative Distribution of wages by gender, Brazil, 1992 and 2005.



Source: Microdata extracted from Pesquisa Nacional por Amostra de

Figure 8: Kernel density estimator of the log-hourly wage PDF by gender, Brazil, 1981 and 1992

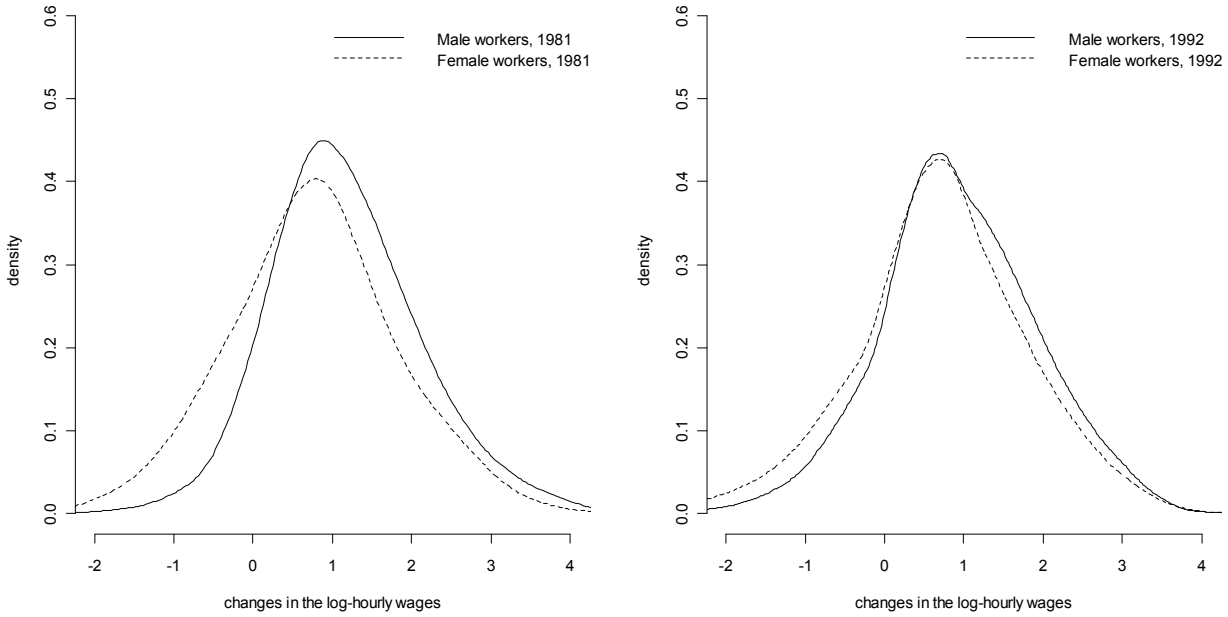
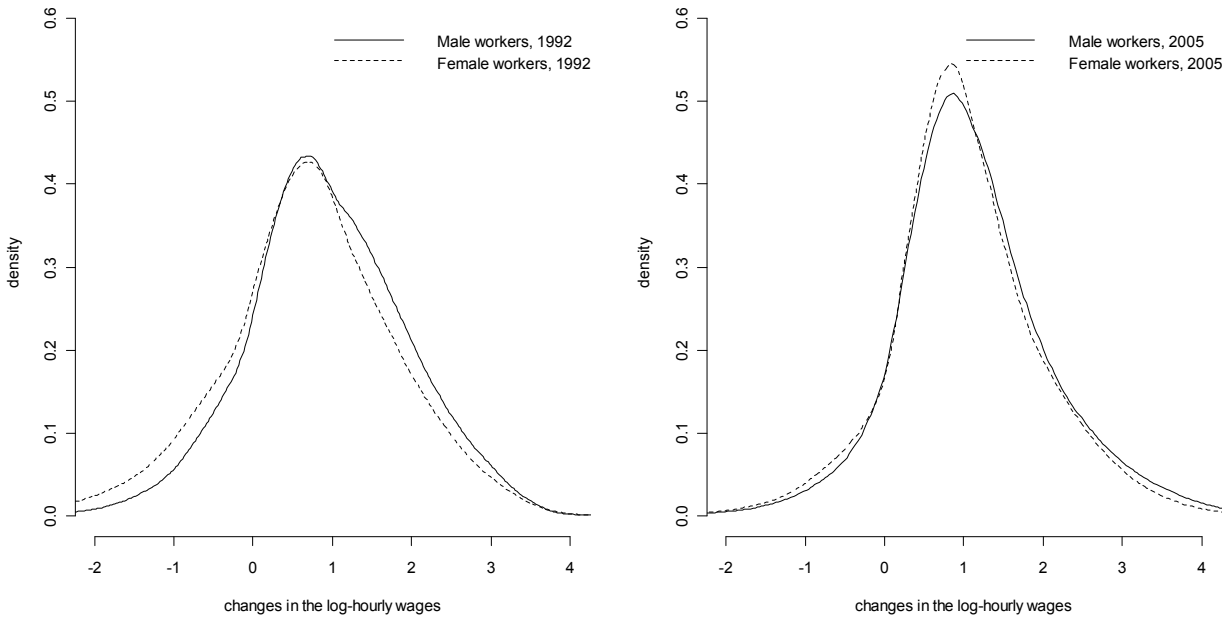


Figure 9: Kernel density estimator of the log-hourly wage PDF by gender, Brazil, 1992 and 2005



Source: Microdata extracted from Pesquisa Nacional por Amostra de