This problem set consists of 4 pages. Try to answer all questions.

All questions should have an answer in plain text – in some cases possibly very short. For many questions it is also relevant to show the R code you have used. Be sure to make the distinction between plain text and code clear, e.g. by using different fonts.

In the grading, most emphasis is put on a correct and clear answer. Some weight is also put on clean and efficient coding.

I. Households and inequality (weight 65%)

In this exercise, we use two datasets with individual income data. Consider first the data set `cps.xls`, which contains an extract of the US Current Population Survey (CPS). The data set includes the following variables:

`year`	Year in which survey was conducted
`ahe`	Average hourly earnings
`bachelor`	Dummy for having at least at BA
`female`	Dummy for female
`age`	Age

1) Import the data set into R. Find the mean hourly earnings for men and women.
2) Assume that everybody works 8 hours a day, 5 days a week for 50 weeks a year. Find the annual income of each individual. Find the relative income rank for each individual (i.e. relative rank 0 is the poorest and relative rank 1 is the richest).
3) Use the individual ranks from the preceding question to assign each individual to a decile.[1] Compute the fraction of women in each decile and show this in a graph. Comment briefly.
4) We want to compute the Lorenz curve for annual income y. If the data are sorted from low to high, the point on the Lorenz curve corresponding to individual `i` can be computed as
$$L_i = \frac{\sum_{j=1}^{i} y_j}{n \times \bar{y}}$$
Here[2] the numerator is the cumulative sum of annual incomes from the first until individual `i` and the denominator the sample size times the average annual income $\bar{y}$. For example, if there were 10 individuals with incomes 1,2,3,…10 so the average is 5, then $L_5 = \frac{1+2+3+4+5}{10\times 5}$. Compute each point on the Lorenz curve and plot it (y-axis) in a diagram against the relative rank (between 0 and 1) of annual income (on the x-axis) together with a 45° line.
5) Consider a simple tax schedule where a flat tax rate $t = 25\%$ is imposed on everybody. Assume that the total tax income is transferred back to the individuals – everybody gets the same amount, equal to average tax payment per person. Compute incomes after taxes and transfers. Compute and draw the Lorenz curve for this income and compare to the Lorenz curve computed above. Present your two Lorentz curves, (1) before and (2) after tax/transfers, in the same panel with different colors with same type of graph as in (4).

[1] A decile is a division of the data into 10 equally sized groups where the first decile consists of the 10th of the data with the lowest annual income, the second decile the next 10th etc.
[2] In the formula, j is an index that runs from individual 1 (the poorest) to individual i (the one under scrutiny).

We now turn to some data based on the distribution of annual earnings in Norway. These can be found in the data file `incomes_couples.csv`. The data consists of data from households where both husband and wife are working. We have the following variables

`id`	Individual id number
`hhid`	Household id number
`inc`	Annual wage earnings
`female`	Logical variable indicating women

6) Import the data set into R. Compute the Gini coefficient for income for the whole sample.[3] Go on to compute the Gini for men and women separately.
NB These are pre-tax incomes so inequality levels are higher than the numbers usually seen in official statistics.

7) We want to construct a data set with household incomes. Match together men and women from the same household based on their `hhid` and compute total household income for each couple. Compute the Gini coefficient for household incomes. Compute the correlation coefficient between the incomes of husbands and wives.

In most societies, the incomes of husbands and wives are positively correlated. This is often referred to as assortative mating. We want to see how the distribution of household incomes had been if couples had been matched randomly and compare the Gini coefficient of household incomes under this assumption compared to the Gini coefficient of observed household incomes.

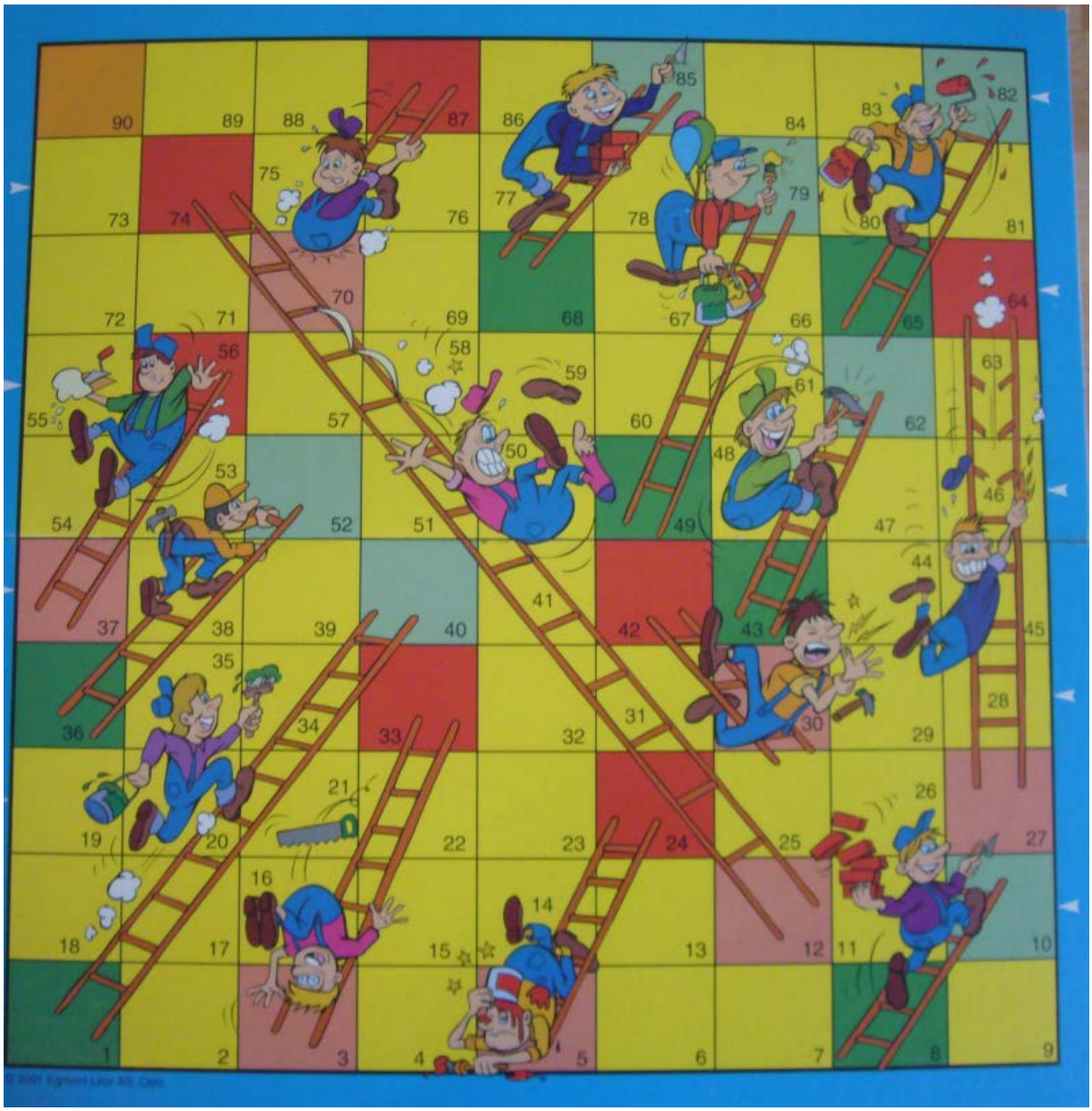
8) Construct a new tibble with men and women matched randomly instead of according to `hhid`. Each man and woman should appear once and only once in the new tibble. Find the household income in the randomly matched couples and compute the Gini coefficient for this hypothetical household income. Verify that the correlation between the incomes of husbands and wives is close to zero in the randomly matched sample.

[3] To compute the Gini coefficient, you can use the `Gini` command from the `ineq` package.

II. Chutes and ladders (weight 35%)

In this exercise, we are going to simulate gameplay in the board game *Chutes and ladders* (also known as *Snakes and ladders*, *Stigespill* in Norwegian). There are many versions of the game. Here we use the Norwegian edition whose board is depicted below. As players do not interact during game play, we will only simulate a single player.

The board consists of 90 squares. You start before square 1. A six-sided die is rolled and you move as many squares as the number of eyes shown on the die. If you end up on a dark green square, you "take the ladder" to the light green square. If you end up on a dark red square, you "fall down the ladder" to the light red square. The game ends when you reach the last square (90). You do not need the exact number of eyes to hit 90, so if you start at square 88 and roll 4 the game ends.



The board has a total of 14 ladders. Their starting and ending squares can be found in the attaches CSV file `ladders.csv`.

1) Consider first a simplified version of the game where we disregard the ladders. Construct a function that makes one simulation of the game and return the number of rolls required to finish the game. Simulate 10 000 rounds of game play. Find the average number of turns it take to complete the game and plot the distribution of the number of turns.
2) Consider now the version depicted in the figure with 14 ladders. Construct a function that makes one simulation of the game and return the number of rolls required to finish the game. Simulate 10 000 rounds of game play. Find the average number of turns it take to complete the game and plot the distribution of the number of turns.
3) Consider next a version of the game where ladders work in both directions, so landing on square 1 takes you to square 40 as before, but now also lading on square 40 takes you to square 1.
   Construct a function that makes one simulation of the game and return the number of rolls required to finish the game. Simulate 10 000 rounds of game play. Find the average number of turns it take to complete the game and plot the distribution of the number of turns. Does this reduce or increase the expected length of gameplay compared to the standard version of the game?
4) Consider finally a giant version of the game with 1000 squares and 200 ladders. The ladders are given in the file `ladders_giant.csv`.
   Construct a function that makes one simulation of the game and return the number of rolls required to finish the game. Simulate 10 000 rounds of game play. Find the average number of turns it take to complete the game and plot the distribution of the number of turns.