Gapminder dataset

Dataset - Gapminder - contains the life expectancy, GDP per capita, and population by country, every five years, from 1952 to 2007. It is an excerpt of a larger and more comprehensive set of data available on Gapminder.org, and the R package of this dataset was created by the statistics professor Jennifer Bryan.

1) Install gapminder - `install.packages("gapminder")`
2) Load gapminder dataset - `data(gapminder)`
3) Return first few lines of dataset
4) Create a vector 'x' of the life expectancies of each country for the year 1952. Plot a histogram of these life expectancies to see the spread of the different countries.

In statistics, the ***empirical cumulative distribution function*** (or empirical cdf or empirical distribution function) is the function *F(a) for any a, which tells you the proportion of the values which are less than or equal to a*.

We can compute F in two ways: the simplest way is to type `mean(x <= a)`. This calculates the number of values in x which are less than or equal a, divided by the total number of values in x, in other words the proportion of values less than or equal to a.

The second way, which is a bit more complex for beginners, is to use the `ecdf()` function. This is a bit complicated because this is a function that doesn't return a value, but a function.

5) What is the proportion of countries in 1952 that have a life expectancy less than or equal to 40?
6) What is the proportion of countries in 1952 that have a life expectancy between 40 and 60 years? [*proportion that have a life expectancy less than or equal to 60 years, minus the proportion that have a life expectancy less than or equal to 40 years*]
7) Plot the empirical cumulative distribution function using `ecdf()`

8) Empirical cumulative distribution using `sapply`

   a) Custom function to ecdf

```
prop <- function(q){
  mean(x <= q)
}
```

   b) Try this out for a value of 'q': `prop(40)`

   c) build a range of q's that we can apply the function to:

```
Qs <- seq(from = min(x), to = max(x), length = 20)
```

   d) use sapply() to apply the 'prop' function to each element of 'qs':

```
props <- sapply(Qs, prop)
```

   e) Take a look at 'props', either by printing to the console, or by plotting it over qs:

```
plot(Qs,props,type="l",
      xlab="a (Life Expectency)",ylab ="Pr(x <= a)")
```

   f) Could be done by anonymous function:

```
props <- sapply(qs, function(q) mean(x <= q))
```

   g) Compare with ecdf.

Normal Distribution

We will use the femaleControlsPopulation.csv. Make sure to put it in your working directory

```
x <- unlist( read.csv("femaleControlsPopulation.csv") )
```

Here x represents the weights for the entire population.

set the seed at 1, then using a for-loop take a random sample of 5 mice 1,000 times. Save these averages. After that, set the seed at 1, then using a for-loop take a random sample of 50 mice 1,000 times. Save these averages.

1) Use a histogram to "look" at the distribution of averages we get with a sample size of 5 and a sample size of 50. How would you say they differ?
2) For the last set of averages, the ones obtained from a sample size of 50, what proportion are between 23 and 25?
3) Now ask the same question of a normal distribution with average 23.9 and standard deviation 0.43.

Population & Samples

We will use the mice_pheno.csv. Make sure to put it in your working directory

```
dat <- unlist( read.csv("mice_pheno.csv") )
```

We will remove the lines that contain missing values:

```
dat <- na.omit( dat )
```

1) Use dplyr to create a vector x with the body weight of all males on the control (chow) diet. What is this population's average?
2) compute the population standard deviation.
3) Set the seed at 1. Take a random sample X of size 25 from x. What is the sample average?
4) Use dplyr to create a vector y with the body weight of all males on the high fat hf) diet. What is this population's average?
5) compute the population standard deviation.
6) Set the seed at 1. Take a random sample Y of size 25 from y. What is the sample average?
7) What is the difference in absolute value between $\overline{y} - \overline{x}$ and $\overline{Y} - \overline{X}$?
8) Repeat the above for females. Make sure to set the seed to 1 before each sample call. What is the difference in absolute value between $\overline{y} - \overline{x}$ and $\overline{Y} - \overline{X}$?
9) For the females, our sample estimates were closer to the population difference than with males. What is a possible explanation for this?