Random Variables and Probability Distributions

Random Variables

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
population <- read.csv("data/femaleControlsPopulation.csv")
population <- unlist(population)

set.seed(1)
sample1 <- sample(population,5)
sample1Mean<-mean(sample1)

set.seed(2)
sample2 <- sample(population,5)
sample2Mean<-mean(sample2)

set.seed(3)
sample3 <- sample(population,5)
sample3Mean<-mean(sample3)

populationMean <- mean(population)
sampleMean <- mean(sample)
```

```
## Warning in mean.default(sample): argument is not numeric or logical: returning
## NA
```

```
meansAbsoluteDifference = abs(populationMean-sampleMean)
cat('Sample 1 mean:',sample1Mean,'\nSample 2 mean:',sample2Mean,'\nSample 3 mean:',sample3Mean)
```

```
## Sample 1 mean: 23.564
## Sample 2 mean: 23.558
## Sample 3 mean: 22.812
```

Here, by running multiple times the cell above, you will notice that the sample mean value will change at each run. At each run, a new sample, of 5 observations, is randomly generated from the same population. The sample mean is then called a 'random variable' because it is not determinist, and depends on a randomly generated sample.

Null distributions

The Female Mice Weights dataset contains the body weight of mices following 2 types of diet: a high fat diet and a more controlled diet. We have an observation here telling us that the average weight of the mices that were on a high fat diet was 3 grams larger than the average of the mice on a controlled diet.

```
femaleMiceWeights <- read.csv("data/femaleMiceWeights.csv")
controlledDiet <- filter(femaleMiceWeights,Diet=="chow") %>% select(Bodyweight) %>% unlist
highfatDiet <- filter(femaleMiceWeights,Diet=="hf") %>% select(Bodyweight) %>% unlist


observation <- mean(highfatDiet) - mean(controlledDiet)
cat("Observation:",observation)
```

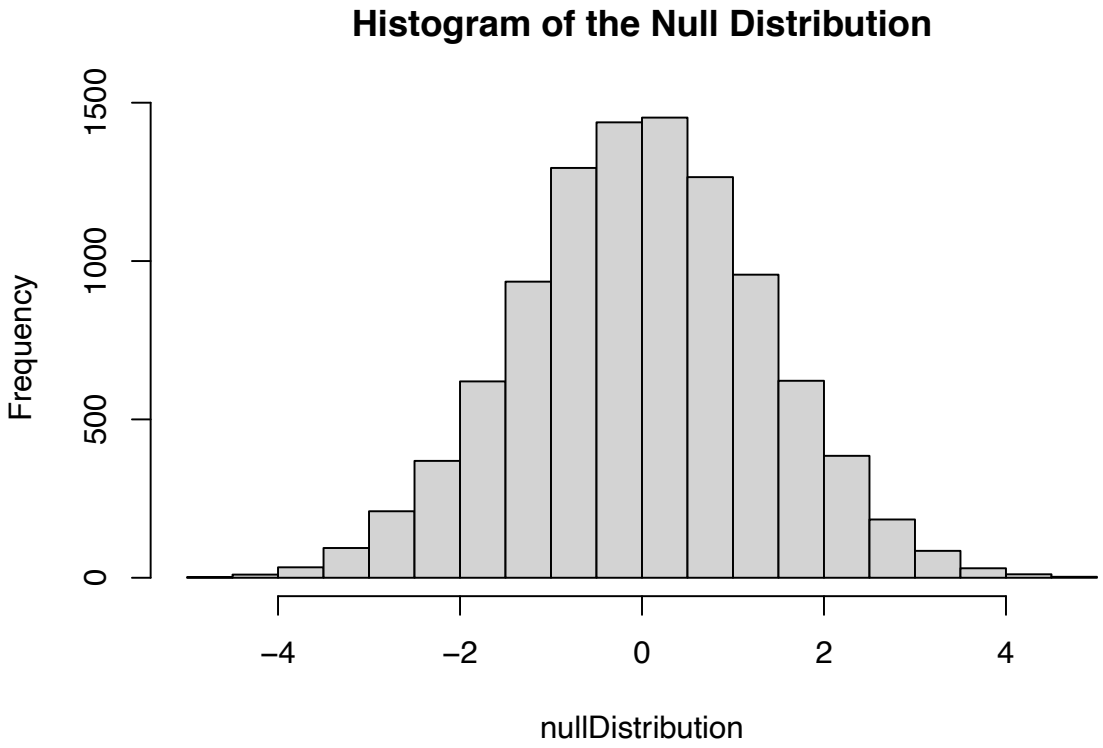
```
## Observation: 3.020833
```

To verify if this observation is reliable, we assume there is no difference between both diets, this is our null hypothesis. Since our null hypothesis says that there is no difference between both diets, we can just take one diet population i.e the controlled diet. The alternative hypothesis says there is a difference between both diet. We can set a significance threshold alpha=0.05 and generate n times, two samples of 12 observations to evaluate our null hypothesis.

We will compute the p-value which is the probability of having our current observations in a world where our null hypothesis is true.

In our case, if we have a p-value $<$ alpha, we will reject the null hypothesis and choose the alternative hypothesis (both diets are different). Subsequently, the lower the p-value, the more meaningful the result because it is less likely to be caused by noise.

```
population <- read.csv("data/femaleControlsPopulation.csv")
population <- unlist(population)

n <- 10000
nullDistribution <- vector("numeric", n)
set.seed(1)
for (i in 1:n){
  controlSample <- sample(population, 12)
  highfatSample <- sample(population, 12)
  nullDistribution[i] <- mean(highfatSample) - mean(controlSample)
}
hist(nullDistribution, main="Histogram of the Null Distribution")
```




```
pvalue <- mean(abs(nullDistribution) > observation)
cat('pvalue:',pvalue)
```

```
## pvalue: 0.0254
```

In this case the p-value is lower than aplha=0.05, thus there is not enough evidence to confirm the null hypothesis. The observations we obtained are very unlikely in a world where the null hypothesis is true. Finally, we can conclude both diets are different.

Probability distributions

The probability distribution is a way of summarizing the observations that we have Knowing the distribution of the probabilities for the observations will help answer questions like : what is the proportion of having someone below a height of 'a' ? $Pr(Height \leq a) = F(a)$

or another question could be : what is the proportion of having someone with a height between 2 numbers 'a' and 'b' ? $Pr(a \leq Height \leq b) = F(b) - F(a)$

We will use the data set called "Gapminder" which is available as an R-package on Github. This data set contains the life expectancy, GDP per capita, and population by country, every five years, from 1952 to 2007.

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 to a, divided by the total number of values in x, in other words the proportion of values less than or equal to a.

```
library(gapminder)
data(gapminder)
head(gapminder)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

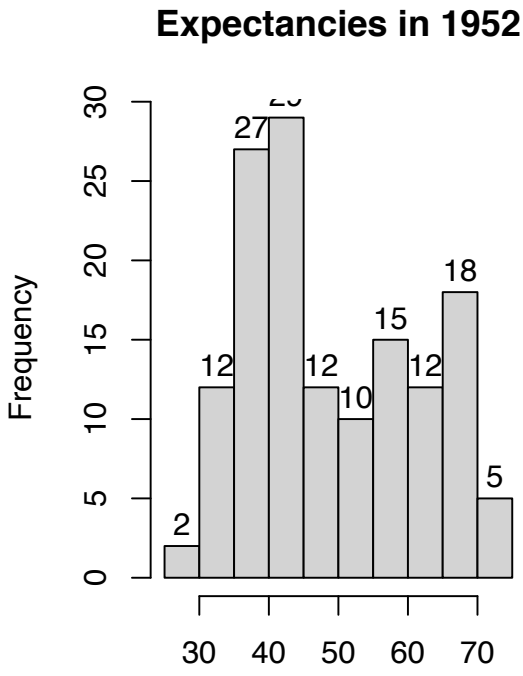
```
library(readr)
library(dplyr)

pop <- unlist(gapminder)

par(mfrow = c(1,2))

#Creates a vector x of the life expectancies of each country for the year 1952.
x <- filter(gapminder,year==1952) %>% select(lifeExp) %>% unlist
prop = function(q) {
  mean(x <= q)
}
cat("Proportion of countries with life expectation lower than 40:",prop(40))
```

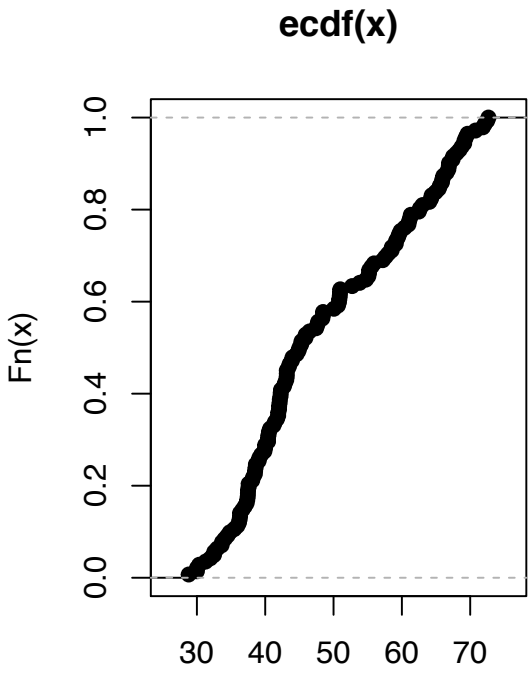
```
## Proportion of countries with life expectation lower than 40: 0.2887324
```

```
cat("Proportion of countries with life expectation between 40 and 60:",prop(60)-prop(40))
```

```
## Proportion of countries with life expectation between 40 and 60: 0.4647887
```

```
hist(x, labels=TRUE, main="Expectancies in 1952")
plot(ecdf(x))
```

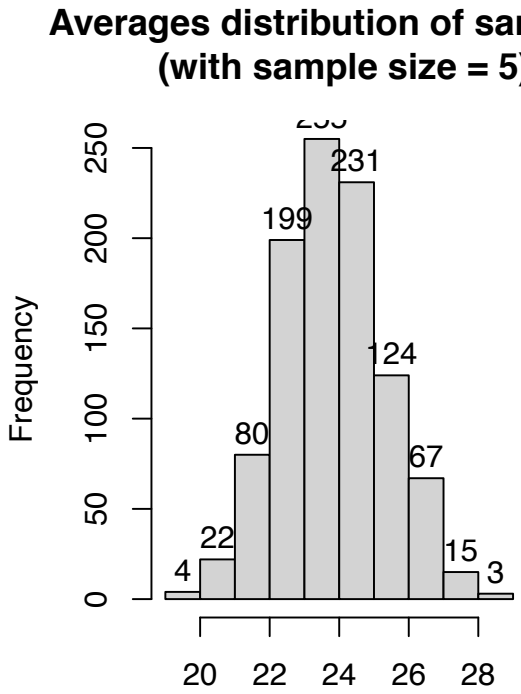


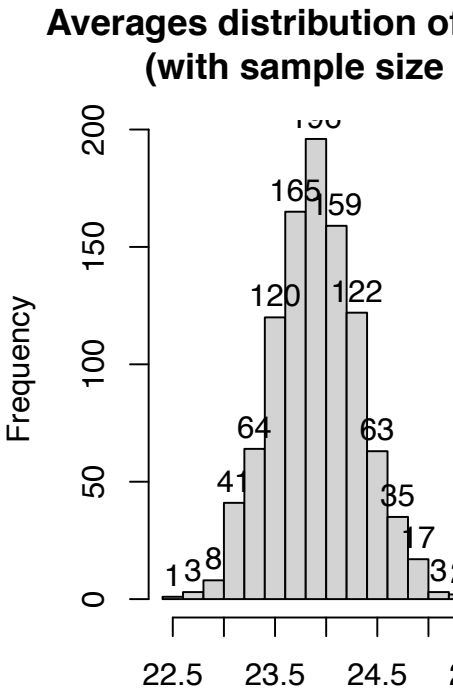


###

The Central Limit Theorem (CLT)

```
population <- read.csv("data/femaleControlsPopulation.csv")
population <- unlist(population)


# make averages5
set.seed(1)
n <- 1000
averages5 <- vector("numeric",n)
for(i in 1:n) {
  X <- sample(population,5)
  averages5[i] <- mean(X)
}

# make averages50
set.seed(1)
n <- 1000
averages50 <- vector("numeric",n)
for(i in 1:n){
  X <- sample(population,50)
  averages50[i] <- mean(X)
}


par(mfrow = c(1,2))
hist(averages5,  labels=TRUE, xlab="Averages ", main="Averages distribution of samples \n (with sample
hist(averages50,  labels=TRUE, xlab="Averages", main="Averages distribution of samples \n (with sample
```





The Normal Distribution

```
cat('Proportion between 23 and 25:',mean( averages50 < 25 & averages50 > 23))
```

```
## Proportion between 23 and 25: 0.982
```

```
mu_averages50 <- mean(averages50)
stdev_averages50 <- sd(averages50)

#We can also use the function pnorm() to find the proportion of observations below a cutoff x given a n
cat('Proportion between 23 and 25:',pnorm(25,mu_averages50,stdev_averages50) - pnorm(23,mu_averages50,s
```

```
## Proportion between 23 and 25: 0.9766827
```

As we can see with these histograms, they both look roughly normal, but with a sample size of 50 the spread is smaller.