Sampling distributions

Chirantan Ganguly

29/04/2020

Sampling From the Population

Lets create a Simple Random Sample of size 100 of birth year from the available population of 21699 baseball players, and find the mean of the sample: -

```
plyr<-read.csv("https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/plyr/baseball.csv"
)
head(plyr)
```

```
##     X        id year stint team lg  g  ab  r  h X2b X3b hr rbi sb cs bb so ibb
## 1   4 ansonca01 1871     1  RC1    25 120 29 39  11   3  0  16  6  2  2  1  NA
## 2  44 forceda01 1871     1  WS3    32 162 45 45   9   4  0  29  8  0  4  0  NA
## 3  68 mathebo01 1871     1  FW1    19  89 15 24   3   1  0  10  2  1  2  0  NA
## 4  99 startjo01 1871     1  NY2    33 161 35 58   5   1  1  34  4  2  3  0  NA
## 5 102 suttoez01 1871     1  CL1    29 128 35 45   3   7  3  23  3  1  1  0  NA
## 6 106 whitede01 1871     1  CL1    29 146 40 47   6   5  1  21  2  2  4  1  NA
##   hbp sh sf gidp
## 1  NA NA NA   NA
## 2  NA NA NA   NA
## 3  NA NA NA   NA
## 4  NA NA NA   NA
## 5  NA NA NA   NA
## 6  NA NA NA   NA
```

```
first_sample<-sample(plyr$year, size=100)
first_sample
```

```
##   [1] 1980 1968 1910 1971 1967 1936 1979 1886 1998 1950 2003 1999 1934 1938 1932
##  [16] 1988 1994 1995 1976 1926 1999 1989 1898 1957 1917 1983 1906 1978 1970 1988
##  [31] 1993 1997 1917 1988 1983 1887 1936 1986 2004 1954 2000 1908 1965 2002 1947
##  [46] 1985 1999 2007 1914 1978 1992 1942 1906 1999 1914 1966 1974 2001 1955 1993
##  [61] 1989 1961 1891 1909 2001 1918 1939 1976 1879 1997 1996 1997 1989 2006 1941
##  [76] 1957 1901 1994 1940 1993 1905 1888 1989 1928 1917 1957 1963 1903 1985 1994
##  [91] 1922 1938 1995 1967 1995 1909 1882 1958 1991 2004
```

```
mean(first_sample)
```

```
## [1] 1959.71
```

Mean of the sampling distribution

The mean of a sample that you take from the population will never be very far away from the population mean (provided that you randomly sample from the population). Furthermore, the mean of the sampling distribution, that is the mean of the mean of all the samples that we took from the population will never be far away from the population mean. Let's observe this in practice.

```
sample_means<-NULL
for(i in 1:500){
  sample_means[i]<-mean(sample(plyr$year, size=200))
}
# The Sampling Mean
mean(sample_means)
```

```
## [1] 1961.044
```

```
#The Population mean
mean(plyr$year)
```

```
## [1] 1961.068
```

Standard deviation of the sampling distribution

In the previous weeks you have become familiar with the concept of standard deviation. You may recall that this concept refers to the spread of a distribution. In R you can calculate the standard deviation using the function sd().

However, the standard deviation of the sampling distribution is called the standard error. The standard error is calculated slightly differently from the standard deviation. The formula for the standard error can be found below:

$s=\sigma/\sqrt{n}$ In this formula, the sigma refers to the standard deviation, while n refers to the sample size of the sample.

```
population_sd<-sd(plyr$year)
population_sd
```

```
## [1] 33.08197
```

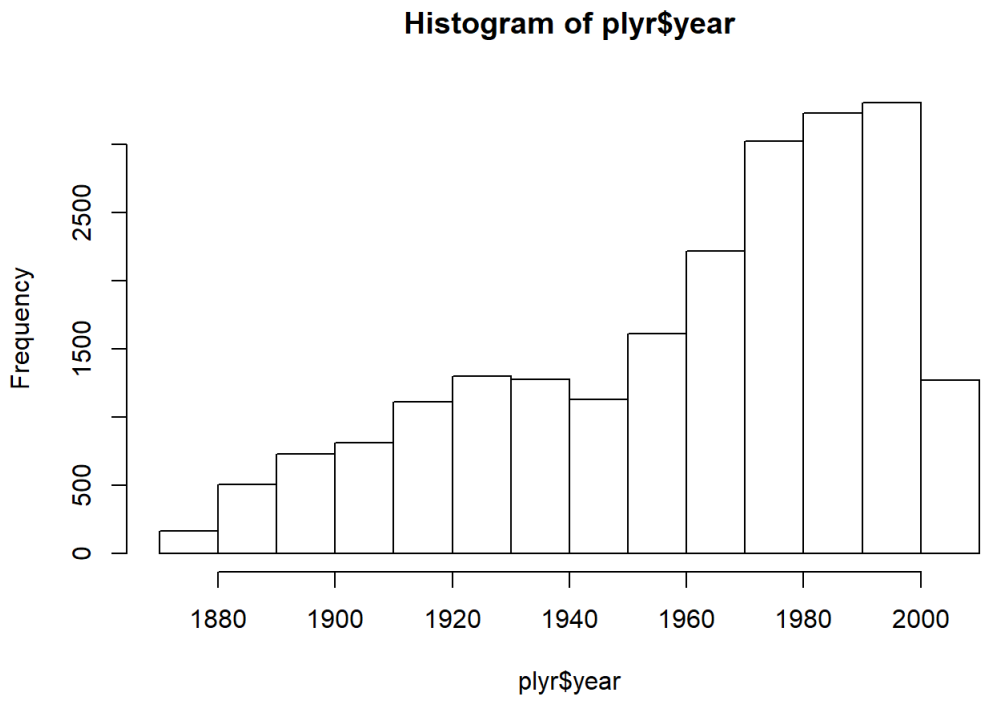
```
sampling_sd<-population_sd/(200^0.5)
sampling_sd
```

```
## [1] 2.339249
```

The central limit theorem

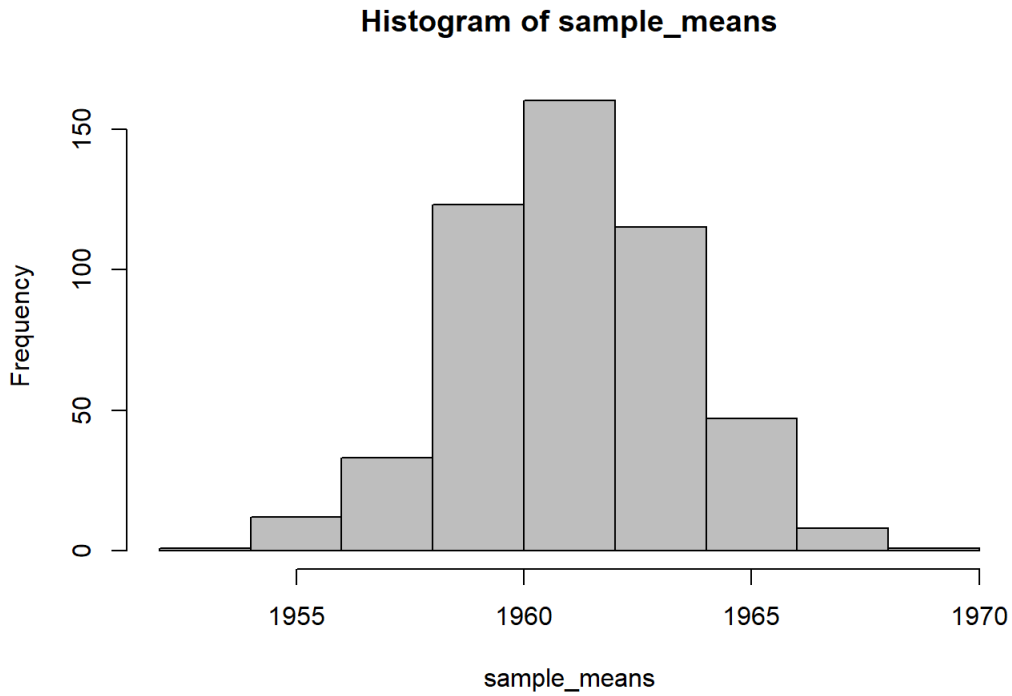
"Provided that the sample size is sufficiently large, the sampling distribution of the sample mean is approximately normally distributed even if the variable of interest is not normally distributed in the population"

```
hist(plyr$year)
```



We can see clearly that the birth year of players is clearly left skewed. However when we plot the sampling mean, we find that its normally distributed: -

```
hist(sample_means, col="grey")
```



Zscores

We already know the concept of z-scores, so lets find the z score of a player born in 1983: -

```
z_score<-(1983-mean(plyr$year))/population_sd
z_score
```

```
## [1] 0.6629656
```

Calculating areas with subjects

In R we can use the pnorm() function to calculate the probability of obtaining a given score or a more extreme score in the population. Basically this calculate an area under the bell curve.

```
pnorm(z_score, lower.tail=TRUE)
```

```
## [1] 0.7463237
```

```
pnorm(z_score, lower.tail=FALSE)
```

```
## [1] 0.2536763
```

Sampling distributions and proportions

The formula of Sample Standard deviation is : - √(pi(1-pi)/n) pi is the sample proportion

```
proportion_hipsters <- 0.10
sample_sd<-(0.1*0.9/200)^0.5
sample_sd
```

```
## [1] 0.0212132
```

```
# calculate the standard deviation of the sampling distribution
sample_sd <- sqrt((0.10 * (1 - 0.10)) / 200)

# calculate the probability
pnorm(0.13, mean = 0.10, sd = sample_sd, lower.tail = FALSE)
```

```
## [1] 0.0786496
```