Week 4 Practice

Jacques Uber

4/22/2021

This is my solution to the second "Analyzing the Databases" question at the end of chapter 7.

```
BLACK_DATABASES = "./data/Databases_in_Excel.xlsx"
df = read_excel(BLACK_DATABASES, sheet = "Consumer Food")
```

The question is: *Consider the Consumer Food database. Compute the mean and standard deviation for the annual food spending for this population. Now take a random sample of 32 of the households in this database and compute the sample mean. Using techniques presented in this chapter, determine the probability of getting a mean that is this large or larger from the population. Work this problem both with and without the finite correction factor and compare the results by discussion the difference in answers.*

First, lets take a look at our data:

```
names(df)
```

```
## [1] "Annual Food Spending ($)"
## [2] "Annual Household Income           ($)"
## [3] "Non mortgage household debt ($)"
## [4] "Region:   1 = NE     2 = MW    3 = S       4 = W"
## [5] "Location:   1 = Metro      2 = Outside Metro"
```

It looks like the first column "Annual Food Spending ($)" is the column of interest. First we go ahead and isolate that columns' data. Next we take a random sample from our `spending` values and compute the variables and statistics the problem asks for.

```
set.seed(123)

n = 32
N = nrow(df)
spending = df$`Annual Food Spending ($)`
rand_spending = sample(spending, n)

mu = mean(spending)
sd = sd(spending)
sample_mu = mean(rand_spending)
sample_sd = sd / sqrt(n)

z = (sample_mu - mu)/sample_sd
z_finite = (sample_mu - mu)/(sample_sd * sqrt((N - n)/(N - 1)))

sprintf("Population size: %s", N)
```

```
## [1] "Population size: 200"
```

```
sprintf("Sample size: %s", n)
```

```
## [1] "Sample size: 32"
```

```
sprintf("Population mean: %s", round(mu, 3))
```

```
## [1] "Population mean: 8966.065"
```

```
sprintf("Population standard deviation: %s", round(sd, 3))
```

```
## [1] "Population standard deviation: 3125.008"
```

```
sprintf("Sample mean: %s", round(sample_mu, 3))
```

```
## [1] "Sample mean: 8065.031"
```

```
sprintf("Sample standard deviation: %s", round(sample_sd, 3))
```

```
## [1] "Sample standard deviation: 552.429"
```

```
sprintf("Infinite-population zscore: %s", round(z, 3))
```

```
## [1] "Infinite-population zscore: -1.631"
```

```
sprintf("Finite-population zscore: %s", round(z_finite, 3))
```

```
## [1] "Finite-population zscore: -1.775"
```

```
sprintf("P(X > z) = %s", round(1 - pnorm(z, 0, 1), 3))
```

```
## [1] "P(X > z) = 0.949"
```

```
sprintf("P(X > z_finite_corrected) = %s", round(1 - pnorm(z_finite, 0, 1),3))
```

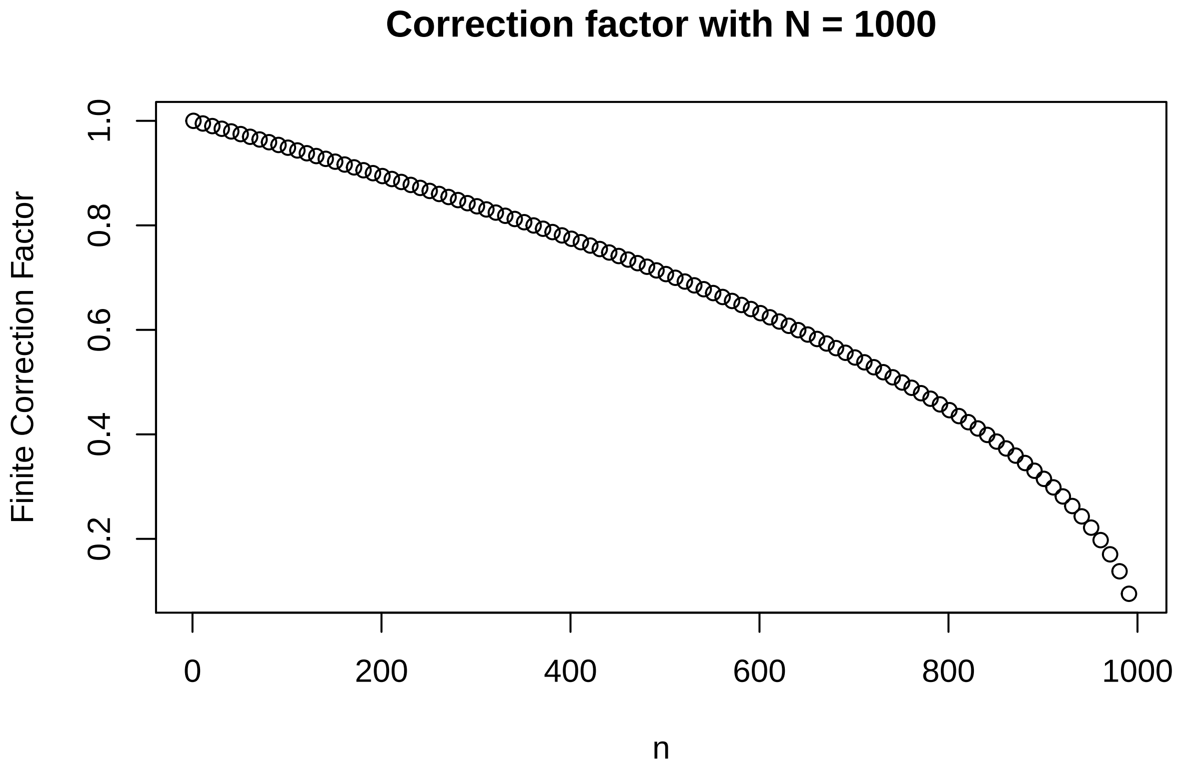
```
## [1] "P(X > z_finite_corrected) = 0.962"
```

Analysis

To analyze these results, I first wanted to see what the finite correction factor ($\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{(\frac{N-n}{N-1})}}$) does as N stays the same and n varies:

```
N = 10^3
n = seq(1, N, 10)

plot(n, sqrt((N - n)/(N - 1)), ylab = "Finite Correction Factor")
title(main=sprintf("Correction factor with N = %s", N))
```



What this chart tells me is that as n approaches N, the factor approaches 0. This means that the denominator of the z score will get smaller and cause the magnitude z score to increase. For the question at hand, we are wondering about the right tail of the normal curve, which represents the probability that a certain mean will be greater than a specific sample mean. When $\bar{x} - \mu$ is negative (and the correction factor is less than 1) the finite z score will be pushed further to the left of a standard zscore making it more likely that a randomly selected sample mean will be greater than it. Similarly, when the $\bar{x} - \mu$ is positive (and the correction factor is less than 1) the finite z score will be pushed further to the right, shrinking the left tail area, making it less likely to find a sample mean that is larger than it.

Our results show that the sample mean (8065.031) was less than the population mean (8966.065). Using the logic in the previous paragraph, we should expect the probability of finding a corrected sample mean to be *greater* than the probability of finding a non-corrected sample mean. Indeed, we see that the left trail area of the corrected sample mean is 0.962 while the uncorrected sample mean is 0.949.