exercise_2_2

Lindsey Greenhill

2/3/2021

Problem 2

Part a

```
births <- tibble(prop_girls =
                   c(.4777,
                     .4875,
                     .4859,
                     .4754,
                     .4874,
                     .4864,
                     .4813,
                     .4787,
                     .4895,
                     .4797,
                     .4876,
                     .4859,
                     .4857,
                     .4907,
                     .5010,
                     .4903,
                     .4860,
                     .4911,
                     .4871,
                     .4725,
                     .4822,
                     .4870,
                     .4823,
                     .4973))
```

```
# calculating standard deviation

sd_actual <- sd(births$prop_girls)

# averaging the proportions above and dividing by the square root of the sample
# size. So p bar divided by the square root of the number. Dividing by 3900
# because we have 3900 observations per month

avg_birth <- mean(births$prop_girls)

std_theoretical <- sqrt(avg_birth * (1 - avg_birth) / 3900)
```

The actual standard deviation in this data set is 0.0064097. The theoretical standard deviation of this data set is 0.0080031. The actual standard deviation is less than the theoretical standard deviation, suggesting that the actual distribution could have a tighter distribution on average than the theoretical distribution if the sexes of the babies were independently decided with a constant probability.

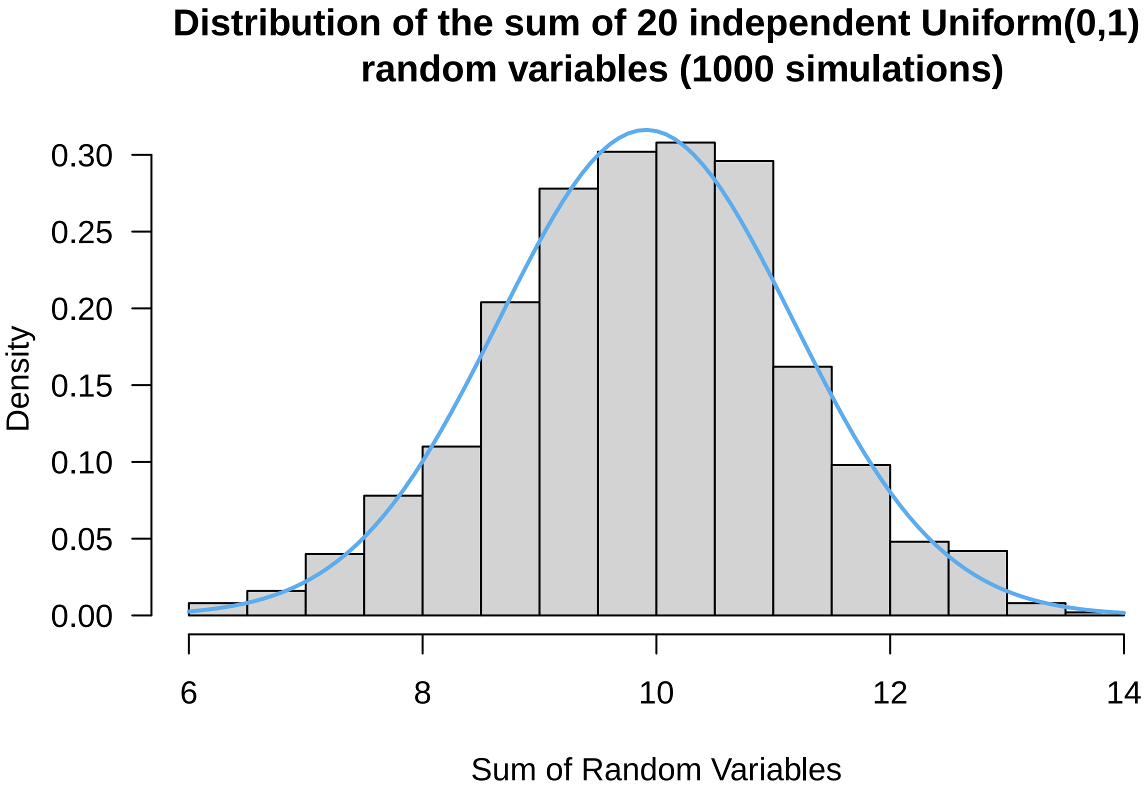
Part b

```r
# use the chi-squared test to do this with 23 df. Finding the upper and lower
# bound and seeing if the chi squared test above is within these bounds.

# finding the chi sq test statistic

chisq <- ((23) * sd_actual^2) / std_theoretical^2

# calculating the upper and lower bounds of the interval

lower_bound <- qchisq(.05/2, df = 23)
upper_bound <- qchisq(.975, df = 23)


# calculating confidence interval

upper_real <- sqrt(23*sd_actual^2 / lower_bound)
lower_real <- sqrt(23*sd_actual^2 / upper_bound)
```

The chi-squared test statistic is **14.7532518**. At the 95% level with 23 degrees of freedom, this value falls within the calculated lower and upper bounds (**11.6885519, 38.0756273**) and is thus not statistically significant. Additionally, the observed standard deviation (0.0064097) is within the calculated range of (0.0049817, 0.0049817).

Problem 3

```r
# creating vector for random draws

sums <- rep(NA, 1000)

for(i in 1:1000){

  # 20 random numbers between 0 and 1

  noms <- runif(20, 0, 1)

  # sum of the random draws

  noms_sum <- sum(noms)

  sums[i] <- noms_sum
}

# plotting a histogram
# getting the mean of the vector

avg_dist <- mean(sums)

{
hist(sums, main = "Distribution of the sum of 20 independent Uniform(0,1)
     random variables (1000 simulations)",
     probability = TRUE, las = 1,
     xlab = "Sum of Random Variables")
curve(dnorm(x=x, mean = mean(sums), sd = sd(sums)),
      col = "steelblue2", lwd = 2, add  = TRUE, yaxt = "n")
}
```



In in a perfect world, the distribution above would be centered at x = 10 and have the same shape and spread as the normal curve. However, because the above distribution is just 1,000 simulations, it does not match the the shape of the normal curve exactly. More specifically, it is not perfectly centered as the normal curve is. However, the mean is very close to 10 (9.911181), demonstrating the power of the central limit theorem.