Central Limit Theorem (CLT)

Main Idea: Sums and averages of random variables from any distribution have approximate normal distributions for sufficiently large sample sizes.

Theorem (Central Limit Theorem)

Suppose $X_1, X_2, \ldots$ *are iid random variables with*

$$E[X_i] = \mu \quad Var[X_i] = \sigma^2.$$

Define

$$\begin{aligned} \text{Sample Sum: } S_n &= X_1 + X_2 + \cdots + X_n \\ \text{Sample Average: } \overline{X}_n &= S_n/n. \end{aligned}$$

Then

$$\begin{aligned} \overline{X}_n &\xrightarrow{d} N(\mu, \sigma^2/n) \\ S_n &\xrightarrow{d} N(n\mu, n\sigma^2) \end{aligned}$$

as $n \to \infty$.

Averages and sums of uniforms

Let $X_i \stackrel{ind}{\sim} Unif(0, 1)$. Then

$$\mu = E[X_i] = \frac{1}{2} \quad \text{and} \quad \sigma^2 = Var[X_i] = \frac{1}{12}.$$

Thus

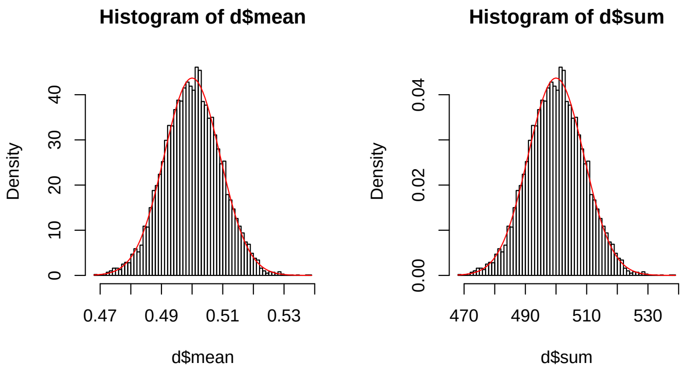
$$\overline{X}_n \sim N\left(\frac{1}{2}, \frac{1}{12n}\right)$$

and

$$S_n \sim N\left(\frac{n}{2}, \frac{n}{12}\right).$$

```
n_sims <- 10000
n_obs  <- 1000
d <- data.frame(rep = rep(1:n_sims, each = n_obs),
                x   = runif(n_sims * n_obs)) %>%
  group_by(rep) %>%
  summarize(mean = mean(x),
            sum  = sum(x))

opar = par(mfrow=c(1,2))
hist(d$mean, 50, probability = TRUE)
curve(dnorm(x, mean = 1/2, sqrt(1/12/n_obs)), add = TRUE, col = "red")
hist(d$sum,  50, probability = TRUE)
curve(dnorm(x, mean = n_obs/2, sqrt(n_obs/12)), add = TRUE, col = "red")
```



Normal approximation to a binomial

Recall that a binomial distribution can be considered as a sum of iid Bernouli random variables, i.e. if $Y = \sum_{i=1}^{n} X_i$ where $X_i \overset{ind}{\sim} Ber(p)$, then

$$Y \sim Bin(n, p).$$

For a binomial random variable, we have

$$E[Y] = np \qquad \text{and} \qquad V[Y] = np(1-p).$$

Now, if n is large,

$$Y \dot{\sim} N(np, np[1-p]).$$

where $\dot{\sim}$ indicates *appoximately distributed*.

Roulette example

A European roulette wheel has 39 slots: one green, 19 black, and 19 red. If I play black everytime, what is the probability that I will have won more than I lost after 99 spins of the wheel?

Let Y indicate the total number of wins and assume $Y \sim Bin(n, p)$ with $n = 99$ and $p = 19/39$. The desired probability is $P(Y \geq 50)$. Then

$$P(Y \geq 50) = 1 - P(Y < 50) = 1 - P(Y \leq 49)$$

```
n = 99
p = 19/39
1-pbinom(49, n, p)
```

```
[1] 0.399048
```

We can approximate Y using $X \sim N(np, np(1-p))$.

$$P(Y \geq 50) \approx 1 - P(X < 50)$$

```
1-pnorm(50, n*p, sqrt(n*p*(1-p)))
```

```
[1] 0.3610155
```

Astronomy example

An astronomer wants to measure the distance, d, from the observatory to a star. The astronomer takes 30 measurements that she believes are unbiased and finds the average of these measurements to be 29.4 parsecs and variance to be 4 parsecs^2. What is the probability the average is within 0.5 parsecs?

Let X_i be the i^{th} measurement. The astronomer assumes that $X_1, X_2, \ldots X_n$ are iid with $E[X_i] = d$ (unbiased) and $Var[X_i] = \sigma^2 = 4$. The estimate of d is

$$\overline{X}_n = \frac{(X_1 + X_2 + \cdots + X_n)}{n} = 29.4.$$

and, by the Central Limit Theorem, we believe $\overline{X}_n \sim N(d, \sigma^2/n)$ where $n = 30$. We want to find

$$\begin{aligned}
P\left(|\overline{X}_n - d| < 0.5\right) &= P\left(-0.5 < \overline{X}_n - d < 0.5\right) \\
&= P\left(\frac{-0.5}{\sigma/\sqrt{n}} < \frac{\overline{X}_n - d}{\sigma/\sqrt{n}} < \frac{0.5}{\sigma/\sqrt{n}}\right) \\
&= P\left(\frac{-0.5}{\sigma/\sqrt{n}} < Z < \frac{0.5}{\sigma/\sqrt{n}}\right) \\
&\approx P(-1.37 < Z < 1.37) \\
&= P(Z < 1.37) - P(Z < -1.37) \\
&\approx 0.915 - 0.085 = 0.830
\end{aligned}$$

Astronomy example (cont.)

Suppose the astronomer wants to be within 0.5 parsecs with 95% probability. How many more samples would she need to take?

We solve

$$\begin{aligned}
0.95 \geq P\left(\left|\overline{X}_n - d\right| < .5\right) &= P\left(-0.5 < \overline{X}_n - d < 0.5\right) \\
&= P\left(\frac{-0.5}{\sigma/\sqrt{n}} < \frac{\overline{X}_n - d}{\sigma/\sqrt{n}} < \frac{0.5}{\sigma/\sqrt{n}}\right) \\
&= P(-z < Z < z) \\
&= 1 - [P(Z < -z) + P(Z > z) \\
&= 1 - 2P(Z < -z)]
\end{aligned}$$

where $z = 0.5/(\sigma/\sqrt{n}) = 1.96$ since

```
-qnorm(.025)

[1] 1.959964
```

and thus $n = 61.47$ which we round up to $n = 62$ to unsure the probability is *at least* 0.95.