§10.5 Central Limit Theorem

Theorem (CLT) Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with finite mean μ and variance σ^2. For each $n \geq 1$, let $S_n = X_1 + \cdots + X_n$. Then

$$\frac{\frac{S_n}{n} - \mu}{\sigma/\sqrt{n}} \text{ converges in distribution to } Z \sim N(0,1) \text{ as } n \to \infty.$$

This means $$\lim_{n\to\infty} P\left(\frac{\frac{S_n}{n} - \mu}{\sigma/\sqrt{n}} \leq t \right) = P(Z \leq t)$$

for any $t \in \mathbb{R}$ (i.e. the CDF of $\frac{\frac{S_n}{n} - \mu}{\sigma/\sqrt{n}}$ converges to the CDF of Z).

Interpretation and intuition

has approximately the same distribution

① $$\frac{\frac{S_n}{n} - \mu}{\sigma/\sqrt{n}} \approx Z \sim N(0,1) \quad \text{for large } n$$

② $$\frac{S_n}{n} \approx \mu + \underbrace{\frac{\sigma}{\sqrt{n}} Z}_{\sim N(0, \frac{\sigma^2}{n})} \quad \text{for large } n$$

SLLN $\frac{S_n}{n} \approx \mu$

CLT $\frac{S_n}{n} \approx \mu + \text{ERROR}$ and the distribution of the ERROR term is $N(0, \frac{\sigma^2}{n})$

③ $$S_n \approx \underbrace{n\mu + \sigma\sqrt{n}\, Z}_{\sim N(n\mu, n\sigma^2)}$$

We previously learned the sum of n i.i.d normals is normal with mean $n\mu$ and variance $n\sigma^2$.

The CLT tells us the same is (approximately) true regardless of the distribution of the terms being summed!

Example A bank teller's service time for each customer is exponentially distributed with mean 2 minutes, independently from customer to customer. Let Y be the total time the teller spends helping 50 customers. Estimate the probability that the teller spends between 90 and 110 minutes.

Let $X_1, \dots, X_{50} \sim Exp(\lambda)$, $\lambda = \frac{1}{2}$, be the i.i.d. service times. Note $\mu = E[X_i] = \frac{1}{\lambda} = 2$

$\sigma^2 = V(X_i) = \frac{1}{\lambda^2} = 4$ and $Y = X_1 + \dots + X_n$, with $n = 50$.

By the CLT, the distribution of Y is approximately

$N(n\mu, n\sigma^2)$ where $n\mu = 50(2) = 100$, $n\sigma^2 = 50(4) = 200$

So $P(90 < Y < 110)$ can be approximated using R and the pnorm command:

```{r}
pnorm(110,100,sqrt(200)) - pnorm(90,100,sqrt(200))
```

```
[1] 0.5204999
```

Example In the previous set-up, estimate the probability that the average service time among the 50 customers is more than 2.5 minutes.

We want to approximate $P\left(\frac{Y}{n} > 2.5\right)$, with $n = 50$.

By the CLT, the distribution of $\frac{Y}{n}$ is approximately the distribution of $\mu + \frac{\sigma}{\sqrt{n}} N(0,1) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

where $\mu = 2$, $\sigma^2 = 4$, and $\frac{\sigma^2}{n} = \frac{4}{50} = 0.08$.

So $P\left(\frac{Y}{n} > 2.5\right)$ can be approximated with the pnorm command:

```{r}
1-pnorm(2.5,2,sqrt(0.08))
```

```
[1] 0.03854994
```

Problem 1. Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with probability mass function

$$P(X_i = k) = \begin{cases} 0.6 & k = +1 \\ 0.4 & k = -1. \end{cases}$$

Think of each X_i as the outcome of one round of a game where you win or lose \$1 with a slight bias to win \$1 on each round. Use the Central Limit Theorem and the `pnorm` command in R to approximate the probability that after 40 rounds of the game your net winnings are between \$4 and \$6.

Let $S = X_1 + \cdots + X_{40}$ be the net winnings after 40 rounds.

By the CLT, the distribution of S is approximately $N(n\mu, n\sigma^2)$

where $n = 40$, $\mu = E[X_i] = (0.6)(1) + (0.4)(-1) = 0.2$, and

$\sigma^2 = E[X_i^2] - \mu^2 = 1 - 0.2^2 = 0.96$. Therefore

$P(4 < S < 6) \approx 0.1141$

Problem 1

```{r}
n = 40; mu = 0.2; sigma = sqrt(0.96)
pnorm(6,n*mu,sqrt(n)*sigma) - pnorm(4,n*mu,sqrt(n)*sigma)
```

```
[1] 0.1141403
```

Problem 2. Consider a continuous distribution with probability density function

$$f(x) = \begin{cases} 3x^2 & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose you go into R and generate 30 random numbers from this distribution, sampling independently. Use the Central Limit Theorem and the `pnorm` command in R to approximate the probability that the sample average your 30 random numbers is in interval $(0.7, 0.8)$.

Let $X_1, \ldots, X_{30}$ be i.i.d with density f. Note

$\mu = E[X_i] = \int_0^1 3x^3\,dx = \frac{3}{4}$, $\sigma^2 = E[X_i^2] - \mu^2 = \int_0^1 3x^4\,dx - \frac{9}{16} = \frac{3}{80}$.

By the CLT, the distribution of $\bar{Y} = \dfrac{X_1 + \cdots + X_{30}}{30}$ is

approximately $N(\mu, \frac{\sigma^2}{30})$ and $P(0.7 < \bar{Y} < 0.8) \approx 0.8427$

Problem 2

```{r}
n = 30; mu = 3/4; sigma = sqrt(3/80)
pnorm(0.8, mu, sigma/sqrt(n)) - pnorm(0.7, mu, sigma/sqrt(n))
```

```
[1] 0.8427008
```

Problem 3. Suppose you have invited 64 guests to a party and need to determine how much food to buy. You believe that each guest will eat 0, 1, or 2 sandwiches with probability 1/6, 1/2, and 1/3 respectively. Assume that the number of sandwiches each guests is independent from other guests.

a. Use the Central Limit Theorem and the **pnorm** command in R to approximate the probability that your guests eat less than 75 sandwiches in total.

b. The 95th percentile of the $N(\mu, \sigma^2)$ distribution is the number $q \in \mathbb{R}$ defined so that if $X \sim N(\mu, \sigma^2)$ then $P(X \leq q) = 0.95$. Within R, you can find the 95th percentile (or other percentiles) using the command **qnorm(0.95, mu, sigma)**. Use this concept to find the fewest number of sandwiches you should buy so that there is at most a 5% chance of having a shortage of sandwiches.

Let $X_1, \ldots, X_{64}$ be the number of sandwiches eaten by each of the 64 guests and let $S = X_1 + \cdots + X_{64}$.

Note $\mu = E[X_i] = 0(1/6) + 1(1/2) + 2(1/3) = 7/6$ and

$\sigma^2 = E[X_i^2] - \mu^2 = 0^2(1/6) + 1^2(1/2) + 2^2(1/3) - \frac{49}{36} = \frac{17}{36}$.

By the CLT, the distribution of S is approximately $N(64\mu, 64\sigma^2)$.

(a) $P(S < 75) \approx 0.52417$

(b) The 95th percentile of $N(64\mu, 64\sigma^2)$ is $q = 83.709$. Therefore $P(S > q) \approx 0.05$ and we should buy at least 84 sandwiches

Problem 3

```{r}
n = 64; mu = 7/6; sigma = sqrt(17/36)
pnorm(75, n*mu, sqrt(n)*sigma) # part a
qnorm(0.95, n*mu, sqrt(n)*sigma)
```

```
[1] 0.5241746
[1] 83.70921
```