Lecture 5: Sampling Distributions and The Central Limit Theorem
STAT 630, Fall 2020

Recall that:

- A **parameter** is a numerical summary of a population (e.g., the population mean μ). It is a fixed number and usually unknown.
- A **statistic** is a numerical summary of a sample (e.g., the sample mean $\bar{x}$). It is random since it varies from sample to sample.

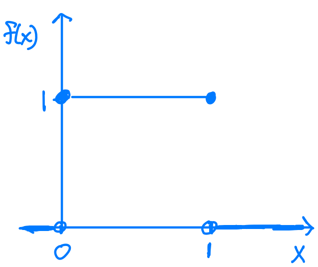
A **sampling distribution** is the distribution of values of a statistic when repeatedly taking random samples of the same size from a population.

Simulation Study
Suppose a population has a uniform distribution between 0 and 1, denoted by $U(0,1)$.

(a) Let X be a random variable such that $X \sim U(0,1)$. Find the mean and variance of X?

The probability density function for $X \sim U(0,1)$ is given by

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$



$$\mu = E(X) = \int_0^1 x f(x) dx = \int_0^1 x dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

To compute the variance use $Var(X) = E(X^2) - [E(X)]^2 = E(X^2) - \mu^2$

$$E(X^2) = \int_0^1 x^2 f(x) dx = \int_0^1 x^2 dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$$

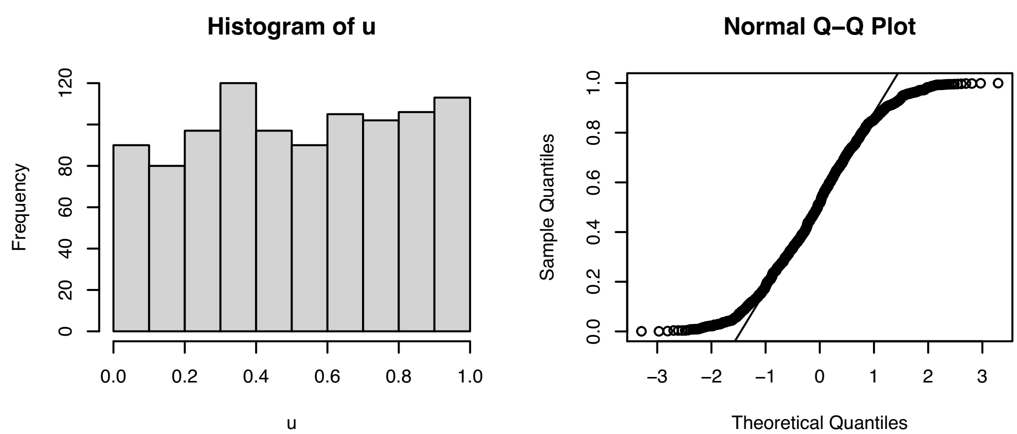
Therefore,

$$\sigma^2 = Var(X) = E(X^2) - \mu^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

(b) Use R to draw 1000 random numbers from $U(0,1)$. Make a histogram and normal QQ plot of the values. Also, compute the mean and variance of the values.

```r
set.seed(100)
u <- runif(1000)

par(mfrow=c(1,2), cex=0.6)
hist(u)
qqnorm(u)
qqline(u)
```




```r
mean(u)

## [1] 0.5180817

var(u) # close to 1/12 = 0.0833

## [1] 0.08254194
```

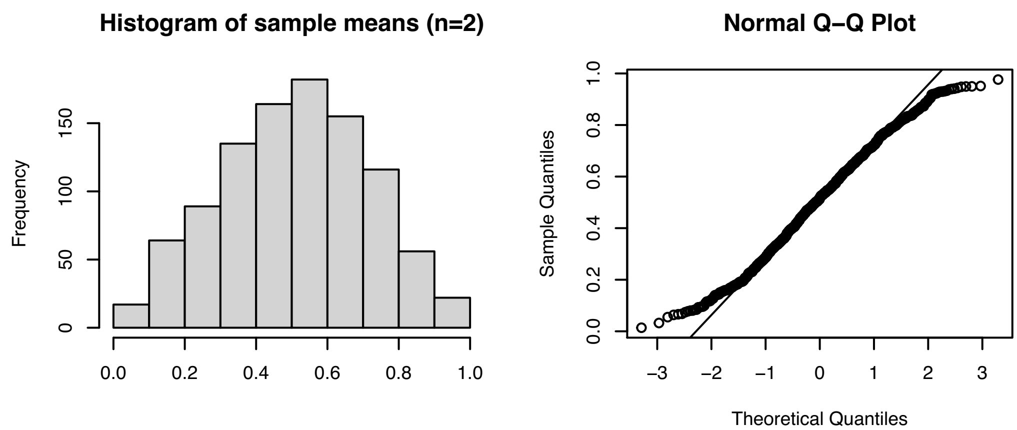
Notice that:

$$\texttt{mean(u)} \approx 0.5 = E(X) = \mu$$
$$\texttt{var(u)} \approx \frac{1}{12} = Var(X) = \sigma^2$$

(c) Use R to repeatedly draw 1000 samples of size $n = 2$ from $U(0, 1)$. Take the sample mean of the values in each sample. Make a histogram and normal QQ plot of the 1000 sample means. Compute the mean and variance of the sample means. What do you notice?

```
set.seed(100)
xbars <- rep(0, 1000) # initialize vector
for(i in 1:1000) {
  samp <- runif(2)
  xbars[i] <- mean(samp)
}

par(mfrow=c(1,2), cex=0.6)
hist(xbars, main = "Histogram of sample means (n=2)", xlab='')
qqnorm(xbars)
qqline(xbars)
```




```
mean(xbars)

## [1] 0.5104061

var(xbars)

## [1] 0.04123252
```

What do you notice?

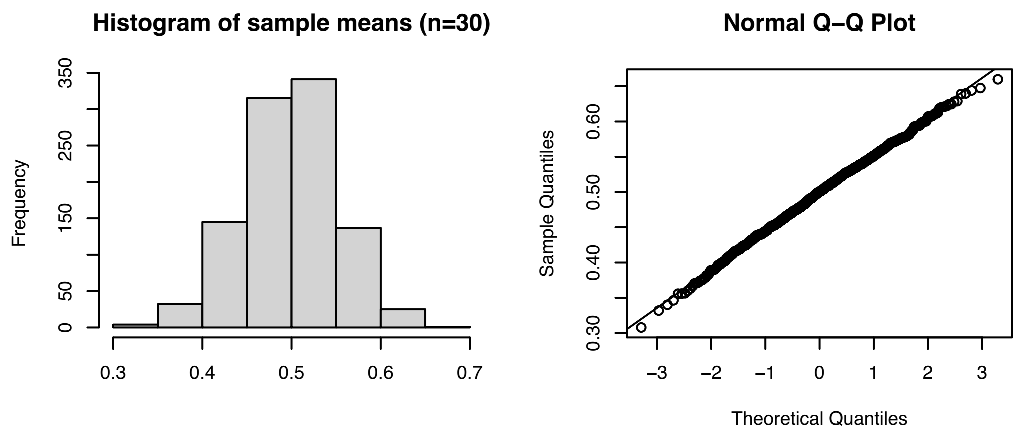
The histogram looks bell-curve shaped. However, the QQ plot has an S-shape, indicating some deviations from the normal distribution (shorter tails, data are less dispersed than a normal distribution).

Also, `mean(xbars)` $\approx 0.5 = \mu$ and `var(xbars)` $\approx \frac{1/12}{2} = \frac{\sigma^2}{2}$

(d) Use R repeatedly draw 1000 samples of size $n = 30$ from $U(0, 1)$. Take the sample mean of the values in each sample. Make a histogram and normal QQ plot of the 1000 sample means. Compute the mean and variance of the sample means. What do you notice?

```
set.seed(100)
xbars <- rep(0, 1000) # initialize vector
for(i in 1:1000) {
  samp <- runif(30)
  xbars[i] <- mean(samp)
}

par(mfrow=c(1,2), cex=0.6)
hist(xbars, main="Histogram of sample means (n=30)", xlab='')
qqnorm(xbars)
qqline(xbars)
```




```
mean(xbars)

## [1] 0.4985099

var(xbars)

## [1] 0.002890106
```

What do you notice?

The histogram and QQ plot indicate that the sample means are normally distributed when the sample size $n = 30$.

Also, `mean(xbars)` $\approx 0.5 = \mu$ and `var(xbars)` $\approx \frac{1/12}{30} = \frac{\sigma^2}{30}$

The Central Limit Theorem

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a population with mean μ and standard deviation σ. Specifically, $X_1, X_2, \cdots, X_n$ are independent and identically distributed (i.i.d.) random variables such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Define the sample mean and total as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$T = \sum_{i=1}^{n} X_i$$

The Central Limit Theorem (CLT) states that when n is large the sample mean $\bar{X}$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. This is true regardless of the shape of the population distribution for X. To summarize, for large n:

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

To transform $\bar{X}$ to a standard normal distribution use:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Similarly, the CLT also states that the sample total $T \sim N(n\mu, \sqrt{n}\sigma)$ for large n.
To transform T to a standard normal distribution use:

$$Z = \frac{T - n\mu}{\sqrt{n}\sigma}$$

Remarks:

- Simulation studies have suggested that $n \geq 30$ is a large enough sample size for the CLT to hold. However, do not apply this rule blindly. For highly skewed populations we might need a sample size larger than 30. For populations that are symmetric, sample sizes smaller than 30 might be sufficient.
- If the population distribution is normal, then $\bar{X}$ is normally distributed for any sample size n.

Ex1. Let X be a random variable with $\mu = 10$ and $\sigma = 4$. A sample of size of 100 is taken from this population.

(a) Find the probability that the sample mean of these 100 observations is less than 9.

Since n is large $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ by the CLT.

So $\bar{X} \sim N(10, 4/\sqrt{100}) = N(10, 0.4)$

$$P(\bar{X} < 9) = P\left(Z < \frac{9 - 10}{0.4}\right) = P(Z < -2.5) = \texttt{pnorm(-2.5)} = \boxed{0.0062}$$

(b) Find the probability that the sum of these 100 observations is greater than 950.

Since n is large $T \sim N(n\mu, \sqrt{n}\sigma)$ by the CLT.

So $T \sim N(100 \cdot 10, \sqrt{100} \cdot 4) = N(1000, 40)$

$$P(T > 950) = 1 - P(T < 950) = 1 - P\left(Z < \frac{950 - 1000}{40}\right)$$
$$= 1 - P(Z < -1.25) = 1 - \texttt{pnorm(-1.25)} = \boxed{0.8943}$$

Ex2. A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean $\mu = 205$ pounds and standard deviation $\sigma = 15$ pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported.

Let $T =$ the total weight of the 49 boxes. Since n is large $T \sim N(n\mu, \sqrt{n}\sigma)$ by the CLT.

So $T \sim N(49 \cdot 205, \sqrt{49} \cdot 15) = N(10045, 105)$

Hence, the probability all boxes can be safely loaded is

$$P(T < 9800) = P\left(Z < \frac{9800 - 10045}{105}\right) = P(Z < -2.33) = \texttt{pnorm(-2.33)} = \boxed{0.0099}$$

Theorem. Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed (i.i.d.) random variables. Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Show that $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2/n$.

To show this use the following properties of expectation of variance. Let X and Y be random variables, and a and b constants.

- $E(aX + b) = aE(X) + b$
- $Var(aX + b) = a^2 Var(X)$
- $E(X + Y) = E(X) + E(Y)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- If X and Y are independent then $Cov(X, Y) = 0$, and so $Var(X + Y) = Var(X) + Var(Y)$

Using these properties:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{\mu} = \mu$$

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \stackrel{\text{indep}}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$