Lecture 5: Sampling Distributions and The Central Limit Theorem
STAT 630, Fall 2021

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.

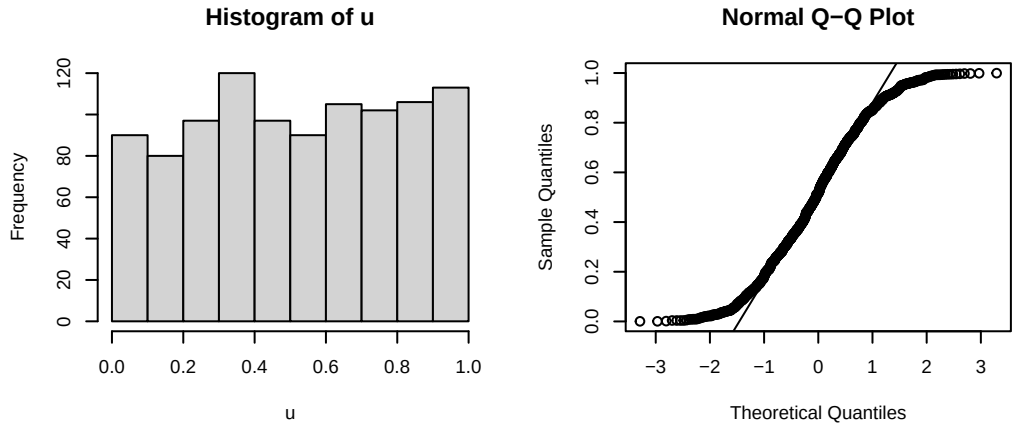
A **sampling distribution** is the distribution of values of a statistic when repeatedly taking random samples of the same size from a population.

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?

(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.

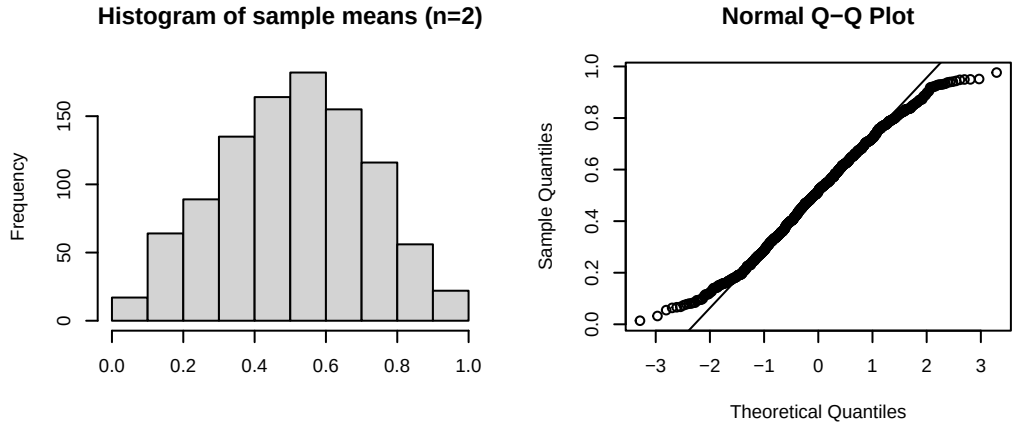
```
set.seed(100)
u <- runif(1000)

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




```
mean(u)

## [1] 0.5180817

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

## [1] 0.08254194
```

(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?

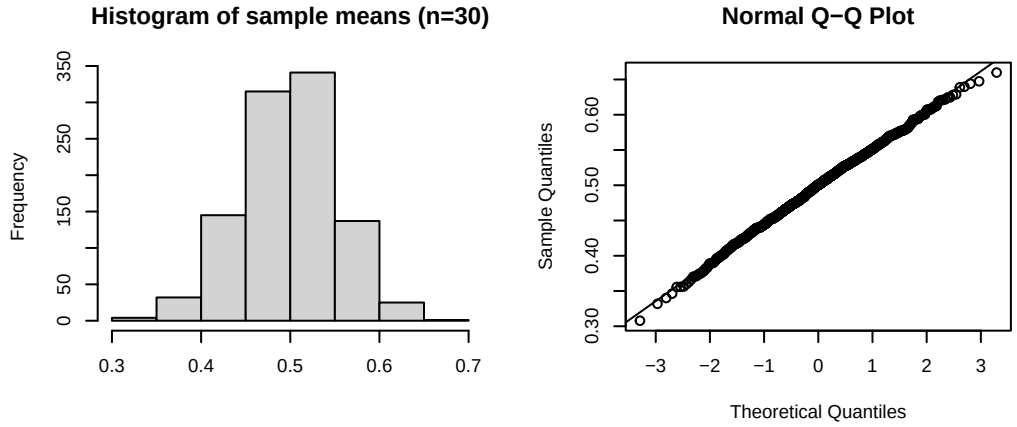
```
set.seed(100)
xbars <- rep(0, 1000) # initialize vector
for(i in 1:1000) {
  samp <- runif(2)
  xbars[i] <- mean(samp)
}

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
```

(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
```

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 $n = 100$ is taken from this population.

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

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

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.

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 $Var(X + Y) = Var(X) + Var(Y)$