# Associate BI Project

## Connor Lough

## 5/2/2021

## Background (Question 2. b)

Related deep dive regarding the correlation between evaluation stage sales opportunities. In our dashboard, we saw that there might be a correlation between duration in eval stage and opportunity outcome. Here we'll set up a test to determine with some confidence that the correlation coefficient is not equal to 0.

H_0: R_lw = 0
H_1: R_lw != 0
setting alpha at 0.05

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Creating a distribution of Point-Biserial correlation coefficients.

```r
names(opps) <- c("oppID","outcome","evalDuration")
r_distr <- 1:1000
set.seed(4)
for (i in 1:length(r_distr)) {
  opps_testing <- sample_n(opps, nrow(opps)*.25)
  opps_lost <- opps_testing[opps_testing$outcome==0,]
  opps_won <- opps_testing[opps_testing$outcome==1,]

  m_1 <- mean(opps_lost$evalDuration)
  m_0 <- mean(opps_won$evalDuration)

  s_n <- sd(opps_testing$evalDuration)
  n_0 <- nrow(opps_lost)
  n_1 <- nrow(opps_won)
  n <- nrow(opps_testing)
```
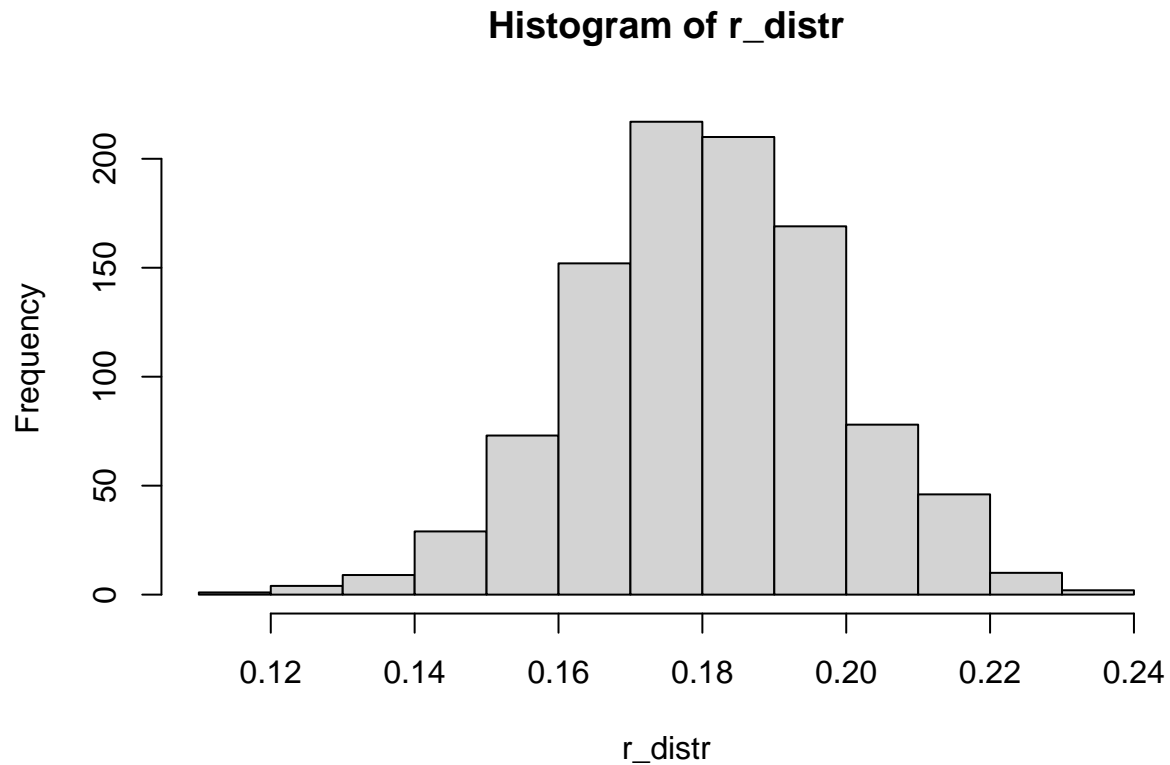
```
  r_distr[i] <- (((m_1)-(m_0))/(s_n))*(sqrt(((n_1)*(n_0))/(n^2)))

}
hist(r_distr)
```

## Histogram of r_distr



Using the central limit theorem, we'll complete our test using the Z statistic. We'll reject H_0 if Z-score is less than, or greater than, -2 & 2 respectively.

```
z_mean <- mean(r_distr)
z_sd <- sd(r_distr)
(0-z_mean)/z_sd
```

```
## [1] -10.08296
```

With a z-score of -10.08, we reject the null hypothesis with 95% certainty.