MSAN 504 - HW3

Andre Guimaraes Duarte

August 1st, 2016

1

We have stock prices from AAPL and SP500 for the years 2002-2006 and 2007-2011 in two files. We wish to test the null hypothesis H_0 that the correlation between the returns in the two stocks is equal for the two periods. In mathematic terms, we have $H_0 : \rho_{20022006} = \rho_{20072011}$. The alternate hypothesis is $H_1 : \rho_{20022006} \neq \rho_{20072011}$, the correlations for the two time periods are not equal. This is a two-sided alternative hypothesis. α is set to 0.05.

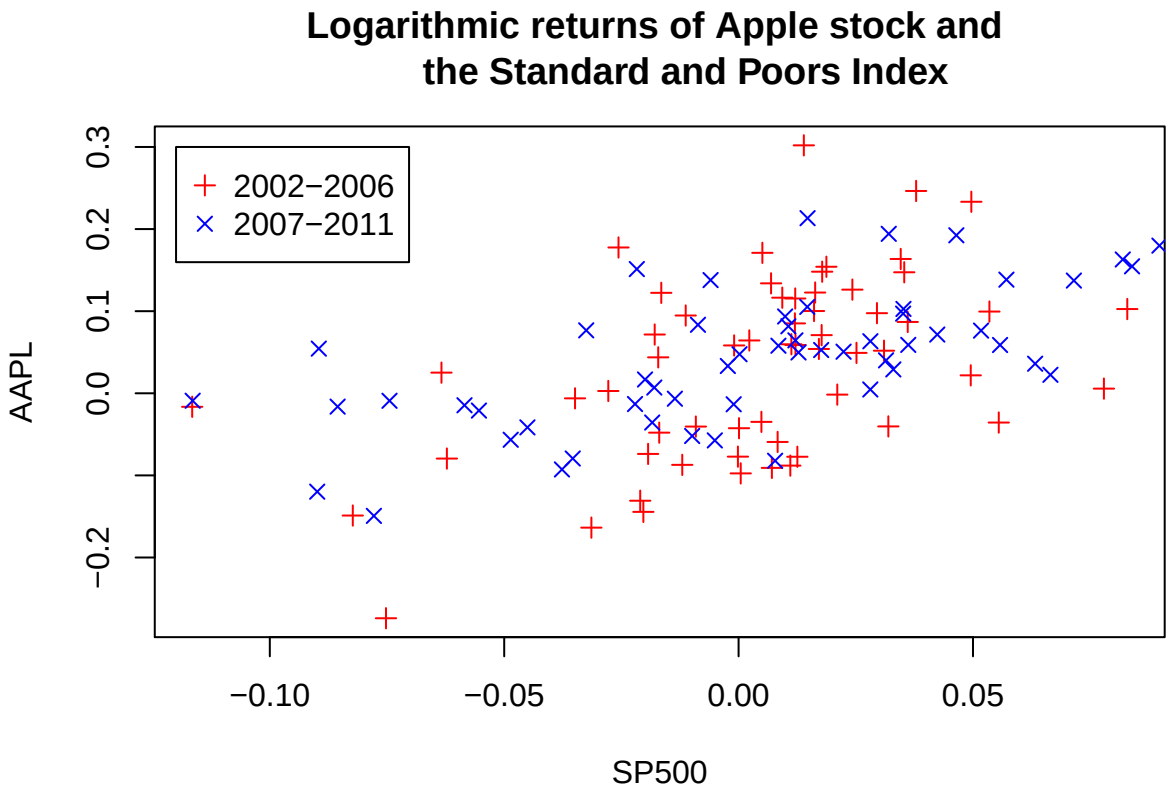
First, we need to conpute the returns on both stocks. We use the logarithmic return formula $\log\left(\frac{p_t}{p_{t-1}}\right)$ that we implemented as a function in R.

```
AAPLSP50020022006 <- read.csv("AAPLSP50020022006.csv", header=T)
AAPLSP50020072011 <- read.csv("AAPLSP50020072011.csv", header=T)

logReturn <- function(vec){
  logRet <- c()
  for(i in seq_along(vec)){
    if(i == 1) next
    logRet[i-1] <- log(vec[i]/vec[i-1])
  }
  return(logRet)
}

AAPL20022006LogReturn <- logReturn(AAPLSP50020022006$AAPL)
SP50020022006LogReturn <- logReturn(AAPLSP50020022006$SP500)
AAPL20072011LogReturn <- logReturn(AAPLSP50020072011$AAPL)
SP50020072011LogReturn <- logReturn(AAPLSP50020072011$SP500)
```

```
plot(AAPL20022006LogReturn~SP50020022006LogReturn, col = "red", pch = 3,
     main="Logarithmic returns of Apple stock and
     the Standard and Poors Index", xlab = "SP500", ylab = "AAPL")
points(AAPL20072011LogReturn~SP50020072011LogReturn, col = "blue", pch = 4)
legend(-0.12, 0.3, c("2002-2006", "2007-2011"), pch=c(3, 4), col=c("red", "blue"))
```



The sample correlation coefficients are found using the formula $\rho_{X,Y} = \frac{cov(X,Y)}{var(X)var(Y)}$, or using R's built-in `cor` function.

```
rho_hat_20022006 <- cor(AAPL20022006LogReturn, SP50020022006LogReturn)
rho_hat_20072011 <- cor(AAPL20072011LogReturn, SP50020072011LogReturn)
```

Here, we get $\widehat{\rho_{20022006}} = 0.471$ and $\widehat{\rho_{20072011}} = 0.608$.

In order to test the hypothesis, we use the formula

$$Z = \frac{F(\widehat{\rho_{20022006}}) - F(\widehat{\rho_{20072011}})}{\sqrt{\frac{1}{n_{20022006}-3} + \frac{1}{n_{20072011}-3}}} \sim N(0,1)$$

where $n_{20022006}$ and $n_{20072011}$ are the sample sizes for each time period, and F is the Fisher transform: $F(x) = \frac{1}{2}\log(\frac{1+x}{1-x})$.

```
fisher <- function(x) .5*log10((1+x)/(1-x))

alpha <- 0.05
n_20022006 <- length(AAPL20022006LogReturn)
n_20072011 <- length(AAPL20072011LogReturn)

F_rho_20022006 <- fisher(rho_hat_20022006)
F_rho_20072011 <- fisher(rho_hat_20072011)

Z <- (F_rho_20022006 - F_rho_20072011) / sqrt(1/(n_20022006 - 3) + 1/(n_20072011 - 3))
p <- pnorm(Z)*2 #two-tailed
```

Here, we have $n_{20022006} = 59$, $n_{20072011} = 59$, $F(\widehat{\rho_{20022006}}) = 0.222$, $F(\widehat{\rho_{20072011}}) = 0.306$, $Z = -0.445$, and $p_{value} = 0.657$.

Since $p_{value} > \alpha$, we do not have enough data at this point in time to reject the null hypothesis that the correlation between Apple stock and the Standard and Poor's Index for the two time periods is equal.

2

We have $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha+1} e^{-\beta x}$ and $g(x) = \lambda e^{-\lambda x}$. So we get

$\frac{f(x)}{g(x)} = \frac{\beta^\alpha}{\lambda\Gamma(\alpha)} e^{(\lambda-\beta)x} x^{\alpha-2}(\alpha - 1 + x(\lambda - \beta))$, so

$\frac{f(x)}{g(x)} = 0 \Leftrightarrow x = 0$ or $x = \frac{1-\alpha}{\lambda-\beta}$.

Since $f(0) = 0$, then we have $c = \frac{f(\frac{1-\alpha}{\lambda-\beta})}{g(\frac{1-\alpha}{\lambda-\beta})} = 1.3601$

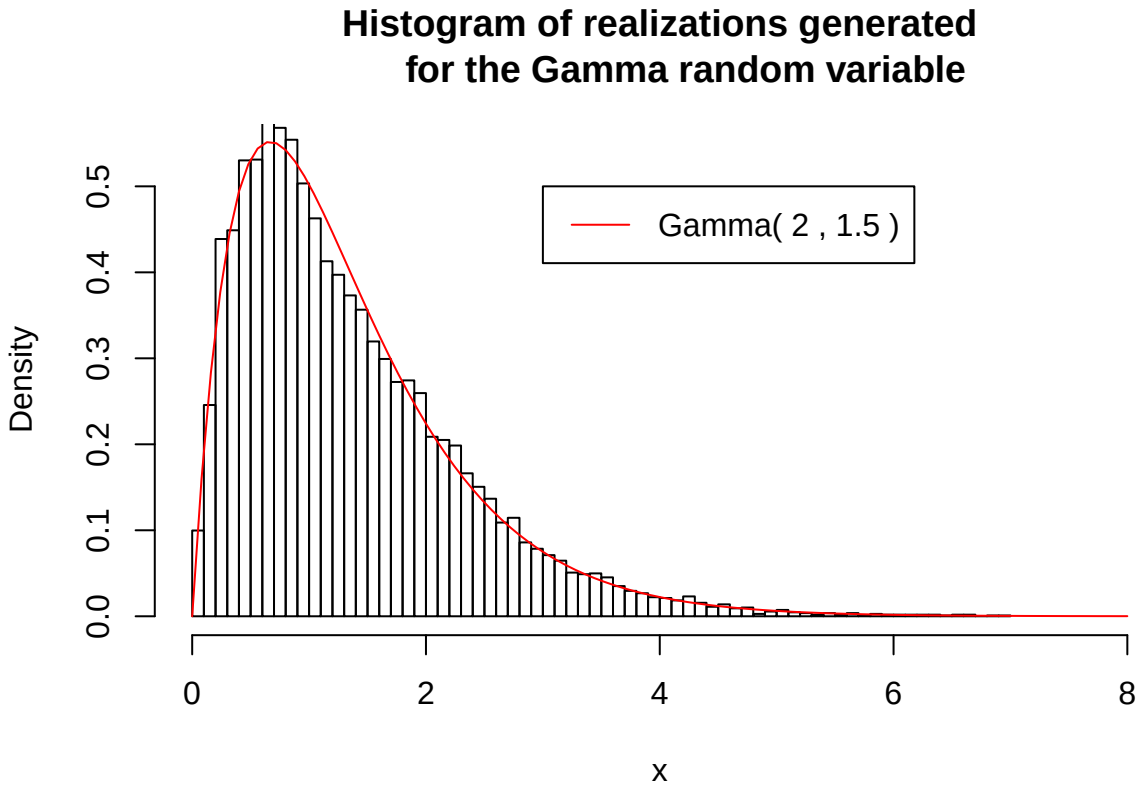
```
alpha <- 2
beta <- 1.5
lambda <- 1/beta

n <- 15000
simulation <- data.frame(unif = runif(n, 0, 1)) # U~unif(0, 1)
simulation$Y <- rexp(n, lambda) # Y~Exp(1/beta)
simulation$auxilliary <- dexp(simulation$Y, lambda) # g(Y)
simulation$target <- dgamma(simulation$Y, alpha, beta) # f(Y)

c <- 1.3601

simulation$accept <- ifelse(simulation$unif < simulation$target/(c*simulation$auxilliary),
                            T, F) # accept if U<f/(c*g)

hist(simulation$Y[simulation$accept], freq=F, breaks=75, ylim=c(0, 0.55), xlim=c(0, 8),
     main="Histogram of realizations generated
     for the Gamma random variable", xlab="x")
curve(dgamma(x, alpha, beta), 0, 8, add=T, col="red")
legend(3, 0.5, paste("Gamma(", alpha, ",", beta, ")"), lty=1, col="red")
```




```
writeLines(paste(sum(simulation$accept), "realizations of the Gamma distribution
out of", n, "realizations of the exponential distribution."))
```

```
## 10826 realizations of the Gamma distribution
## out of 15000 realizations of the exponential distribution.
```

3

We have $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $g(x) = \frac{1}{\pi} \frac{1}{1+x^2}$. So we get

$\frac{f(x)}{g(x)} = \sqrt{\frac{\pi}{2}}(1+x^2)e^{-\frac{x^2}{2}}$, so

$(\frac{f(x)}{g(x)})' = \sqrt{\frac{\pi}{2}}(1+x^2)xe^{-\frac{x^2}{2}}$, so

$(\frac{f(x)}{g(x)})' = 0 \Leftrightarrow (1+x^2) = 0$ or $xe^{-\frac{x^2}{2}} = 0 \Leftrightarrow x = 0$. Therefore,

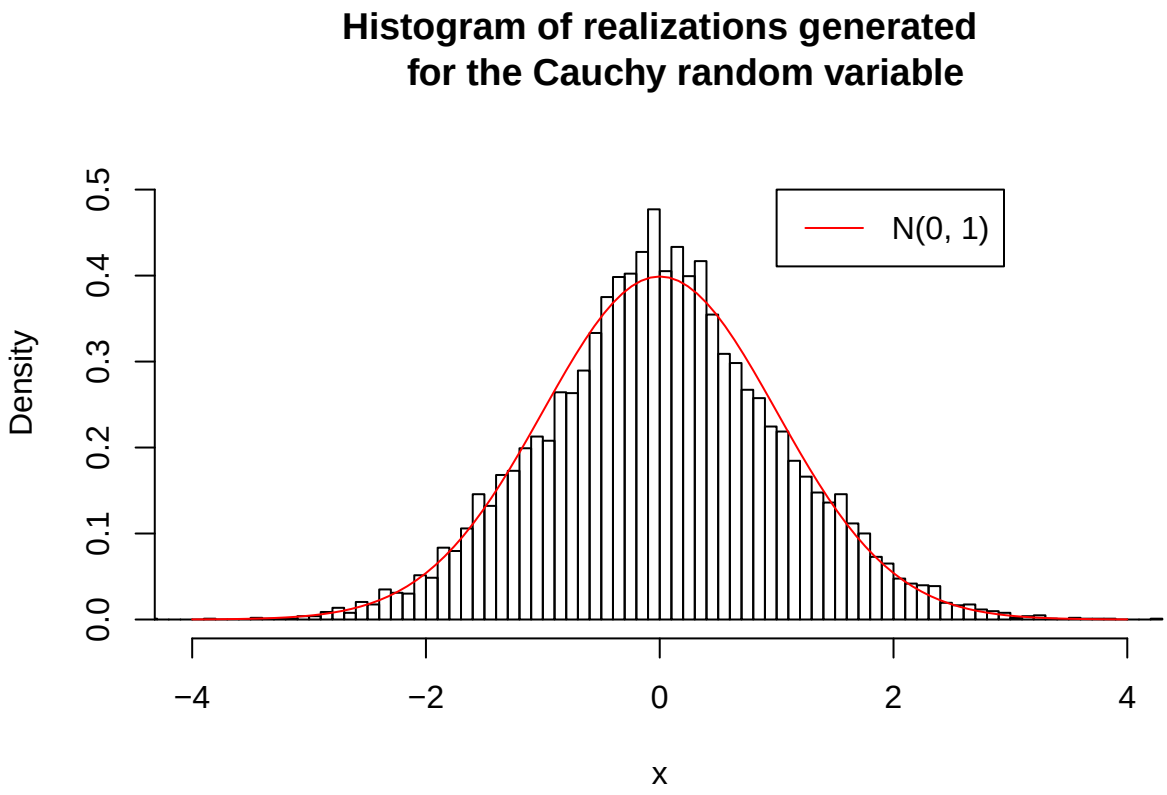
$c = \frac{f(0)}{g(0)} = \sqrt{\frac{\pi}{2}}$.

```
n <- 14000
simulation <- data.frame(unif = runif(n, 0, 1)) # U~unif(0, 1)
simulation$Y <- rcauchy(n) # Y~Cauchy
simulation$auxilliary <- dcauchy(simulation$Y) # g(Y)
simulation$target <- dnorm(simulation$Y, 0, 1) # f(Y)
```

```r
c <- sqrt(pi/2)

simulation$accept <- ifelse(simulation$unif < simulation$target/(c*simulation$auxilliary),
                            T, F) # accept if U<f/(c*g)

hist(simulation$Y[simulation$accept], freq=F, ylim=c(0, 0.55),breaks=75, xlim=c(-4, 4),
     main="Histogram of realizations generated
     for the Cauchy random variable", xlab="x")
curve(dnorm(x, 0, 1), -4, 4, add=T, col="red")
legend(1, 0.5, "N(0, 1)", lty=1, col="red")
```




```r
writeLines(paste(sum(simulation$accept), "realizations of the standard normal distribution
out of", n, "realizations of the Cauchy distribution."))
```

```
## 10292 realizations of the standard normal distribution
## out of 14000 realizations of the Cauchy distribution.
```

4

We want to test the null hypothesis H_0 that men and women have the same average body temperature. This is equivalent to testing whether the difference in their body temperatures is equal to zero. The alternate hypothesis H_1 is that the mean temperature for the two sexes are different, or that the difference in the means is not null. We have no prior knowledge whether men's body temperature is higher or lower than women's, so we decide to do a two-sided test.

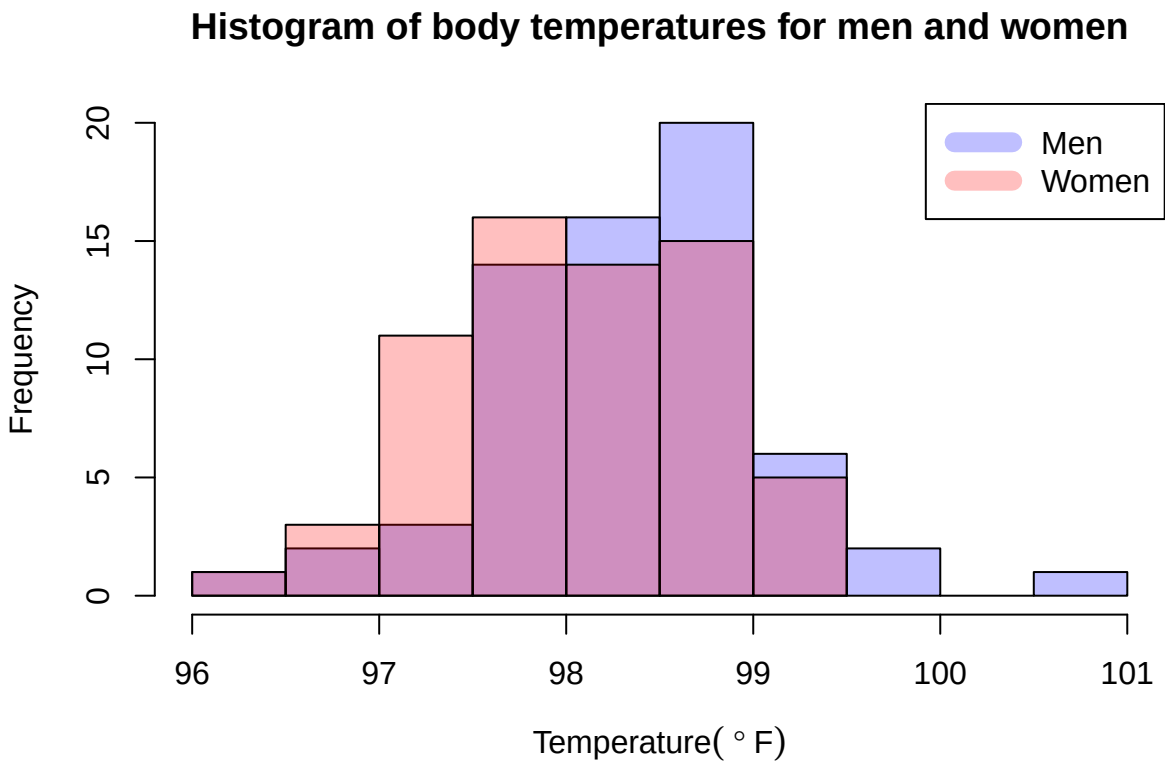
$H_0 : \mu_m = \mu_w \Leftrightarrow \mu_m - \mu_w = 0$

$H_1 : \mu_m \neq \mu_w \Leftrightarrow \mu_m - \mu_w \neq 0$

```
genderTemp <- read.csv("EffectOfGenderBodyTemperaturesAndRestingHeartRate.csv", header=T)
tempMen <- genderTemp$BodyTemp[genderTemp$Gender == 0]
tempWomen<- genderTemp$BodyTemp[genderTemp$Gender == 1]
```

Here we assume that, in the dataset, entries where `gender` is `0` are men, and `1` are women for ease of use in the rest of the problem. But it could just as well be otherwise.

```
plot(h1, col=rgb(0,0,1,1/4), xlim=c(96, 101),
     main="Histogram of body temperatures for men and women",
     xlab = expression(Temperature (~degree~F)))
plot(h2, col=rgb(1,0,0,1/4), xlim=c(96, 101), add=T)
legend("topright", c("Men", "Women"), col=c(rgb(0,0,1,1/4), rgb(1,0,0,1/4)), lwd=10)
```



First, however, we need to check whether the variances in the two groups are the same. We therefore have another null hypothesis H_{0_σ} that the variance in body temperature in men and in women are equal (their ratio is equal to 1), which we will test against H_{1_σ} that the true ratio of variances is not 1. Again, it's a two-sided test because we have no a priori knowledge of the distributions.

The test statistic to use in this case is

$F = \frac{s_m^2}{s_w^2} \sim F(n_m - 1, n_w - 1)$, where s_m^2 and s_w^2 are the sample variances for men and women respectively. We compare this statistic to a Fisher distribution with $n_m - 1$ and $n_w - 1$ degrees of freedom, where n_m and n_w are the sample sizes for men and women respectively.

```
alpha <- 0.05
s2_m <- var(tempMen)
```

```
s2_w <- var(tempWomen)
n_m <- length(tempMen)
n_w <- length(tempWomen)

f <- s2_m/s2_w
p <- pf(f, n_m-1, n_w-1, lower.tail = F)*2 #two-sided
```

Here, we have $s_m^2 = 0.553$, $s_w^2 = 0.488$, $n_m = 65$, $n_w = 65$, $F = 1.132$, and $p_{value} = 0.621$.

Since $p_{value} > \alpha$, we don't have enough data at this point in time to reject the null hypothesis that the variances in the two populations are the same.

We have therefore established that $\sigma_m^2 = \sigma_w^2$. In order to test our initial hypothesis, we use the test statistic $T = \frac{(\bar{x_m} - \bar{x_w}) - (\mu_m - \mu_w)}{s_p\sqrt{\frac{1}{n_m} + \frac{1}{n_w}}} = \frac{\bar{x_m} - \bar{x_w}}{s_p\sqrt{\frac{1}{n_m} + \frac{1}{n_w}}} \sim t(n_m + n_w - 2)$ under H_0, where $s_p = \sqrt{\frac{(n_m-1)s_m^2 + (n_w-1)s_w^2}{n_m+n_w-2}}$ is the pooled standard deviation.

We test this statistic against a Student's t-distribution with $n_m + n_w - 2$ degrees of freedom.

```
xbar_m <- mean(tempMen)
xbar_w <- mean(tempWomen)
s_p <- sqrt(((n_m - 1)*s2_m + (n_w - 1)*s2_w)/(n_m + n_w - 2))

t <- (xbar_m - xbar_w)/(s_p*sqrt(1/n_m + 1/n_w))
p <- pt(t, n_m + n_w - 2, lower.tail = F)*2 #two-sided
```

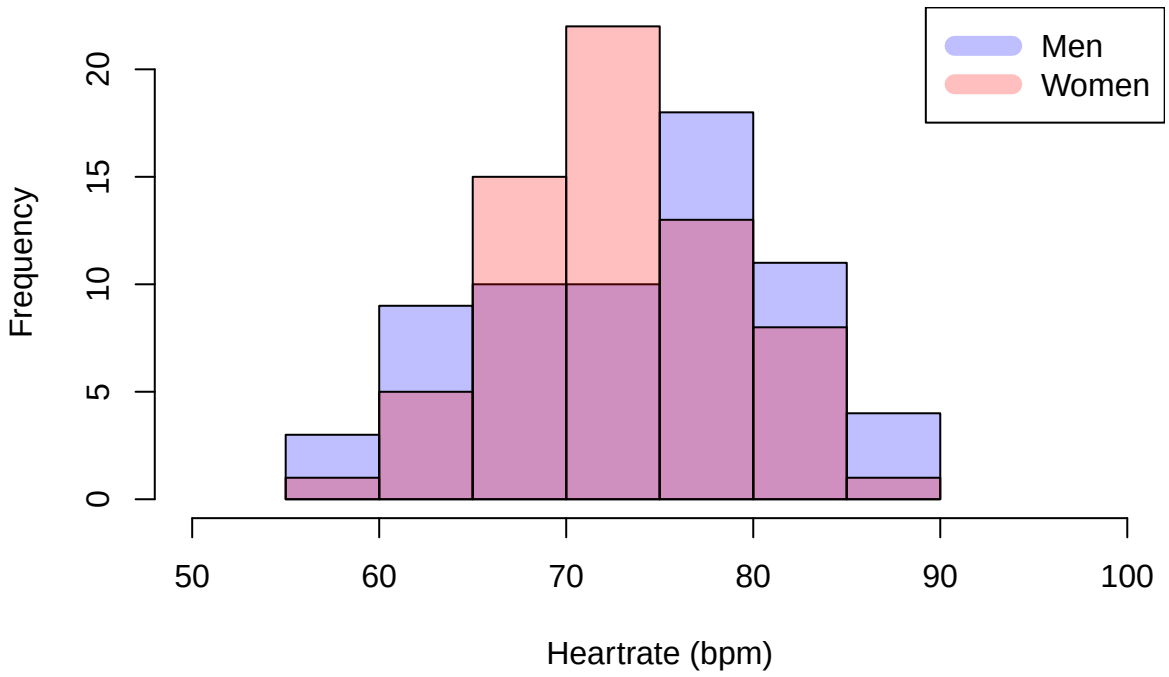
Here, we have $\bar{x_m} = 98.394$, $\bar{x_w} = 98.105$, $s_p = 0.721$, $T = 2.285$, and $p_{value} = 0.024$.

Since $p_{value} < \alpha$, we reject H_0 and conclude that the true difference in body temperature between men and women is not null: men and women have different mean body temperatures.

5

```
genderHeartrate <- read.csv("EffectOfGenderBodyTemperaturesAndRestingHeartRate.csv",
                            header=T)
heartrateMen <- genderHeartrate$Heart.Rate[genderHeartrate$Gender == 0]
heartrateWomen<- genderHeartrate$Heart.Rate[genderHeartrate$Gender == 1]
```

```
plot(h1, col=rgb(0,0,1,1/4), ylim=c(0,22), xlim=c(50,100),
     main="Histogram of heart rates for men and women",
     xlab = "Heartrate (bpm)")
plot(h2, col=rgb(1,0,0,1/4), add=T)
legend("topright", c("Men", "Women"), col=c(rgb(0,0,1,1/4), rgb(1,0,0,1/4)), lwd=10)
```



First, however, we need to check whether the variances in the two groups are the same. We therefore have another null hypothesis H_{0_σ} that the variance in heart rate in men and in women are equal (their ratio is equal to 1), which we will test against H_{1_σ} that the true ratio of variances is not 1. Again, it's a two-sided test because we have no a priori knowledge of the distributions.

The test statistic to use in this case is

$F = \frac{s_m^2}{s_w^2} \sim F(n_m - 1, n_w - 1)$, where s_m^2 and s_w^2 are the sample variances for men and women respectively. We compare this statistic to a Fisher distribution with $n_m - 1$ and $n_w - 1$ degrees of freedom, where n_m and n_w are the sample sizes for men and women respectively.

```
alpha <- 0.1
s2_m <- var(heartrateMen)
s2_w <- var(heartrateWomen)
n_m <- length(heartrateMen)
n_w <- length(heartrateWomen)

f <- s2_m/s2_w
p <- pf(f, n_m-1, n_w-1, lower.tail = F)*2 #two-sided
```

Here, we have $s_m^2 = 65.695$, $s_w^2 = 34.518$, $n_m = 65$, $n_w = 65$, $F = 1.903$, and $p_{value} = 0.011$.

Since $p_{value} < \alpha$, we reject the null hypothesis that the variances are equal.

In this scenario, the 90% confidence interval is found using the formula

$(\bar{x_m} - \bar{x_w}) \pm t^*_{\frac{\alpha}{2}, Satt} \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$, where $\bar{x_m}$ is the sample mean for men, $\bar{x_w}$ is the sample mean for women, n_m is the sample size for men, n_w is the sample size for women, s_m^2 is the variance for men, and s_w^2 is the variance for women.

```
alpha <- 0.1

xbar_m <- mean(heartrateMen)
xbar_w <- mean(heartrateWomen)

Satt <- ((s2_m/n_m + s2_w/n_w)^2) / ((1/(n_m-1))*(s2_m/n_m)^2 + (1/(n_w-1))*(s2_w/n_w)^2)
t_star <- qt(alpha/2, Satt, lower.tail = F)

lower <- (xbar_m - xbar_w) - t_star*sqrt(s2_m/n_m + s2_w/n_w)
upper <- (xbar_m - xbar_w) + t_star*sqrt(s2_m/n_m + s2_w/n_w)
```

Here, we have $t^*_{\frac{\alpha}{2},n_m+n_w-2} = 1.658$.

We get the 90% confidence interval $(-1.274; 2.843)$, and we have $(\bar{x_m} - \bar{x_w}) = 0.785 \in (-1.274; 2.843)$.

The confidence interval for this particular sample is (-1.274; 2.843). Similarly constructed intervals, computed over many different random samples, will contain the true difference of means for heart rate in men and women with probability 90%.

6

By doing a "quick-and-dirty" check, we get $\frac{s_1^2}{s_2^2} = 2.5186477 > 2$. We should then use a formal test to check the null hypothesis H_0 that the true variances between the two groups are equal. The alternate hypothesis is H_1: the true variances are not equal. It is a two-sided test. α is set to 0.01.

The test statistic to use in this case is

$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$, where s_1^2 and s_2^2 are the sample variances for lean and obese volunteers respectively. We compare this statistic to a Fisher distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, where n_1 and n_2 are the sample sizes for lean and obese volunteers respectively.

```
alpha <- 0.01
s2_1 <- 107.121^2
s2_2 <- 67.498^2
n_1 <- 10
n_2 <- 10

f <- s2_1/s2_2
p <- pf(f, n_1-1, n_2-1, lower.tail = F)*2 #two-sided
```

Here, we have $s_1^2 = 1.1474909 \times 10^4$, $s_2^2 = 4555.98$, $n_1 = 10$, $n_2 = 10$, $F = 2.519$, and $p_{value} = 0.185$.

Since $p_{value} > \alpha$, we don't have enough data at this point in time to reject the null hypothesis that the variances in the two populations are the same.

Now that we are in the scenario $\sigma_1 = \sigma_2$, we will be testing the null hypothesis H_0 that lean and obese folks spend the same number of minutes standing or walking ($\mu_1 = \mu_2$). The alternate hypothesis H_1 is that the two groups do not spend the same number of minutes standing or walking ($\mu_1 \neq \mu_2$). This is a two-sided test, and $\alpha = 0.01$.

We use the test statistic

$T = \frac{(\bar{x_1}-\bar{x_2})-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\bar{x_1}-\bar{x_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$ under H_0, where $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ is the pooled standard deviation.

We test this statistic against a Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

```
xbar_1 <- 525.751
xbar_2 <- 373.269
s_p <- sqrt(((n_1 - 1)*s2_1 + (n_2 - 1)*s2_2)/(n_1 + n_2 - 2))

t <- (xbar_1 - xbar_2)/(s_p*sqrt(1/n_1 + 1/n_2))
p <- pt(t, n_1 + n_2 - 2, lower.tail = F)*2 #two-sided
```

Here, we have $\bar{x_1} = 525.751$, $\bar{x_2} = 373.269$, $s_p = 89.529$, $T = 3.808$, and $p_{value} = 0.001$.

Since $p_{value} < \alpha$, we reject H_0 and conclude that the true difference in number of minutes spent standing or walking is not null: lean and obese folks do not spend the same amount of time standing or walking at the $\alpha = 0.01$ level.

7

From problem #6, we had already calculated the p_{value} for the hypothesis test of H_{0_σ} that the variances for group 1 and group 2 are equal (their ratio is equal to 1), which we tested against H_{1_σ} that the true ratio of variances is not 1.

The p_{value} we obtained in that test was $0.185 > \alpha = 0.05$. Therefore, we do not have enough information at this point to reject the null hypothesis that the true variances are equal.

In this scenario, the 95% confidence interval is found using the formula

$(\bar{x_1} - \bar{x_2}) \pm t^*_{\frac{\alpha}{2},n_1+n_2-2} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, where $\bar{x_1}$ is the sample mean for lean folks, $\bar{x_2}$ is the sample mean for obese folks, n_1 is the sample size for lean people, n_2 is the sample size for obese people, and s_p is the pooled standard deviation (detailed formula in #6).

```
alpha <- 0.05
t_star <- qt(alpha/2, n_1+n_2-2, lower.tail = F)

lower <- (xbar_1 - xbar_2) - t_star*s_p*sqrt(1/n_1 + 1/n_2)
upper <- (xbar_1 - xbar_2) + t_star*s_p*sqrt(1/n_1 + 1/n_2)
```

Here, we have $t^*_{\frac{\alpha}{2},n_1+n_2-2} = 2.101$.

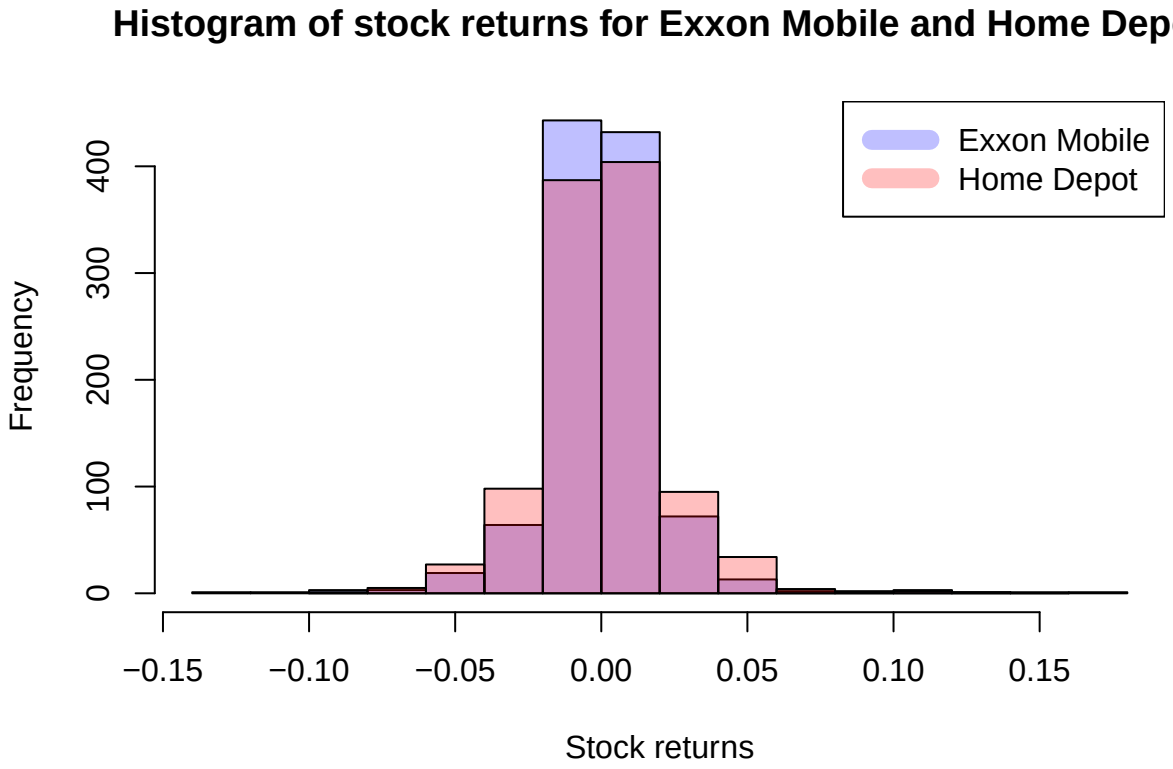
We get the 95% confidence interval $(68.364; 236.6)$, and we have $(\bar{x_1} - \bar{x_2}) = 152.482 \in (68.364; 236.6)$.

The confidence interval for this particular sample is (68.364; 236.6). Similarly constructed intervals, computed over many different random samples, will contain the true difference in the number of minutes spent standing/walking by lean and obese people with probability 95%.

8

We are testing the null hypothesis H_0: the variances for daily stock returns of Exxon Mobil and Home Depot are equal (their ratio is 1), against the alternate hypothesis H_1: the true variances are not equal (their ratio doesn't equal 1). It is a two-sided test, and $\alpha = 0.1$.

```
plot(h1, col=rgb(0,0,1,1/4),
     main="Histogram of stock returns for Exxon Mobile and Home Depot",
     xlab = "Stock returns")
plot(h2, col=rgb(1,0,0,1/4), add=T)
legend("topright", c("Exxon Mobile", "Home Depot"),
       col=c(rgb(0,0,1,1/4), rgb(1,0,0,1/4)), lwd=10)
```



The test statistic to use in this case is

$F = \frac{s_e^2}{s_h^2} \sim F(n_e - 1, n_h - 1)$, where s_e^2 and s_h^2 are the sample variances for the returns of Exxon Mobil and Home Depot respectively. We compare this statistic to a Fisher distribution with $n_e - 1$ and $n_h - 1$ degrees of freedom, where n_e and n_h are the sample sizes for Exxon Mobil and Home Depot respectively.

```
returns <- read.csv("DailyReturnsForFiveStocks.csv", header=T)
alpha <- 0.1
s2_e <- var(returns$XOM)
s2_h <- var(returns$HD)
n_e <- length(returns$XOM)
n_h <- length(returns$HD)

f <- s2_e/s2_h
p <- pf(f, n_e-1, n_h-1)*2 #two-sided
```

Here, we have $s_e^2 = 3.953 \times 10^{-4}$, $s_h^2 = 4.594 \times 10^{-4}$, $n_e = 1060$, $n_h = 1060$, $F = 0.861$, and $p_{value} = 0.015$.

Since $p_{value} < \alpha$, we reject the null hypothesis that the variances for the returns of Exxon Mobil and Home Depot are equal and we accept he alternate hypothesis that the variances are different.

9

We want to compute the 95% confidence interval for the true ratio of the variance of Exxon Mobil daily stock returns to the variance of Home Depot daily stock returns. In order to do so, we use the formula

$P(\frac{1}{F^*_{n_e-1,n_h-1,1-\frac{\alpha}{2}}} \frac{s_e^2}{s_h^2} \leq \frac{\sigma_e^2}{\sigma_h^2} \leq \frac{1}{F^*_{n_e-1,n_h-1,\frac{\alpha}{2}}} \frac{s_e^2}{s_h^2}) = 1 - \alpha$, where n_e is the sample size for Exxon Mobil, n_h is the sample size for Home Depot, s_e^2 is the sample variance for Exxon Mobil, s_h^2 is the sample variance for Home Depot, σ_e^2 is the true variance for Exxon Mobil, and σ_h^2 is the true variance for Home Depot.

```
alpha <- 0.05
F_star_left <- qf(1-alpha/2, n_e-1, n_h-1)
F_star_right <- qf(alpha/2, n_e-1, n_h-1)

lower <- (s2_e/s2_h)/F_star_left
upper <- (s2_e/s2_h)/F_star_right
```

Here, we have $F^*_{n_e-1,n_h-1,1-\frac{\alpha}{2}} = 0.763$ and $F^*_{n_e-1,n_h-1,\frac{\alpha}{2}} = 0.971$.

We get the 95% confidence interval $(0.763; 0.971)$, and we have $\frac{s_e^2}{s_h^2} = 0.861 \in (0.763; 0.971)$.

The confidence interval for this particular sample is $(0.763; 0.971)$. Similarly-constructed intervals, computed over many different random samples, will contain the true ratio of the variance of Exxon Mobil daily stock returns to the variance of Home Depot stock returns $\frac{\sigma_e^2}{\sigma_h^2}$ with probability 95%.

10

We want to test the null hypothesis H_0 that there is no difference in yawning rates between those who are exposed to a "yawn seed" and those who receive no such stimulus. This is equivalent to testing the proportion of yawns in those who have been exposed and those who have not been exposed is equal. The alternate hypothesis H_1 is that the yawning rate is higher in those who have been exposed to a "yawning seed" (the proportion of yawns in those who have been exposed to a "yawning seed" is higher than in those who have not). We do a one-sided test, since we have *a priori* knowledge that "yawns are contagious". α is set to 0.02

$H_0 : \pi_s = \pi_n \Leftrightarrow \pi_s - \pi_n = 0$

$H_1 : \pi_s > \pi_n \Leftrightarrow \pi_s - \pi_n > 0$

where the subscript $_s$ relates to being exposed to a "seed", and the subscript $_n$ relates to not being exposed to such stimulus.

```
yawn <- read.csv("EffectOfSeeingAYawnOnYawning.csv", header=T)

seededYawned <- sum(yawn$Seeded == "Yawn Seeded" & yawn$Yawn == "Yawned")
seededNotYawned <- sum(yawn$Seeded == "Yawn Seeded" & yawn$Yawn == "Didn't Yawn")
notSeededYawned <- sum(yawn$Seeded == "Not Seeded" & yawn$Yawn == "Yawned")
notSeededNotYawned <- sum(yawn$Seeded == "Not Seeded" & yawn$Yawn == "Didn't Yawn")

n_s <- sum(seededYawned, seededNotYawned)
n_n <- sum(notSeededYawned, notSeededNotYawned)

pi_s <- seededYawned/n_s
pi_n <- notSeededYawned/n_n
```

Here, we have $\widehat{\pi_s} = 0.294$, $\widehat{\pi_n} = 0.25$, $n_s = 34$, and $n_n = 16$.

We get $n_s\widehat{\pi_s} = 10$, $n_s(1-\widehat{\pi_s}) = 24$, $n_n\widehat{\pi_n} = 4$, and $n_n(1-\widehat{\pi_n}) = 12$. They are all positive. Therefore, given H_0, we use the test statistic

$Z = \frac{\widehat{\pi_s} - \widehat{\pi_n}}{\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n_s} + \frac{\widehat{\pi}(1-\widehat{\pi})}{n_n}}} \sim N(0,1)$, where $\widehat{\pi}$ is the pooled proportion of yawns.

```
alpha <- 0.02
pi_p <- (sum(seededYawned, notSeededYawned))/(n_s + n_n)

Z <- (pi_s - pi_n)/sqrt((pi_p*(1 - pi_p))/n_s + (pi_p*(1 - pi_p))/n_n)
p <- pnorm(Z, lower.tail = F) #one-sided
```

Here, we have $\pi_p = 0.28$, $Z = 0.324$, and $p_{value} = 0.373$.

Since $p_{value} > \alpha$, we don't have enough data at this point in time to reject the null hypothesis that there is no difference in yawning rate between those who are exposed to a "yawn seed" and those who receive no such stimulus.

11

We want to compute the 80% confidence interval for the true difference in yawning rates if subjects do and do not receive the "yawning seed". In order to do so, we use the formula

$$P((\widehat{\pi_s} - \widehat{\pi_n}) - Z^*_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{\pi_s}(1-\widehat{\pi_s})}{n_s} + \frac{\widehat{\pi_n}(1-\widehat{\pi_n})}{n_n}} \leq (\pi_s - \pi_n) \leq (\widehat{\pi_s} - \widehat{\pi_n}) + Z^*_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{\pi_s}(1-\widehat{\pi_s})}{n_s} + \frac{\widehat{\pi_n}(1-\widehat{\pi_n})}{n_n}}) = 1 - \alpha$$

where the subscript $_s$ relates to being exposed to a "seed", and the subscript $_n$ relates to not being exposed to such stimulus..

```
alpha <- 0.2
Z_star <- qnorm(alpha/2, lower.tail = F)

lower <- (pi_s - pi_n) - Z_star*sqrt((pi_s*(1-pi_s))/n_s + (pi_n*(1-pi_n))/n_n)
upper <- (pi_s - pi_n) + Z_star*sqrt((pi_s*(1-pi_s))/n_s + (pi_n*(1-pi_n))/n_n)
```

Here, we have $Z^*_{\frac{\alpha}{2}} = 1.282$.

We get the 80% confidence interval (-0.127; 0.215), and we have $(\widehat{\pi_s} - \widehat{\pi_n}) \in (-0.127; 0.215)$.

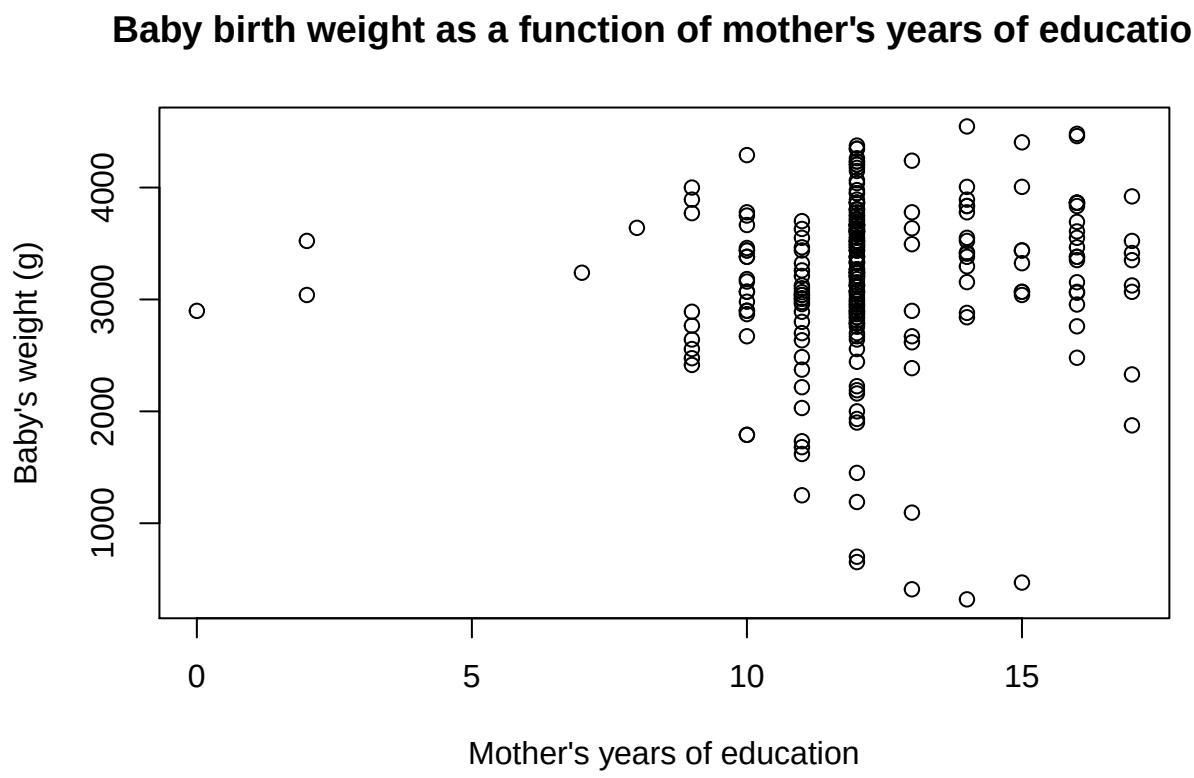
The confidence interval for this particular sample is (-0.127; 0.215). Similarly-constructed intervals, computed over many different random samples, will contain the true ratio of the difference in yawning rate if subjects do or do not receive the "yawning seed" $\frac{\pi_s}{\pi_n}$ with probability 80%.

12

a

We want to see whether better-educated mothers tend to have heavier (and therefore possibly healthier) babies?

```
myData <- read.csv("MotherEducationBirthWeight.csv", header=T)
plot(myData$BIRTHWEIGHT~myData$YEARSEDUC,
     main="Baby birth weight as a function of mother's years of education",
     xlab="Mother's years of education",
     ylab="Baby's weight (g)")
```




```
rho_hat_0 <- cor(myData$YEARSEDUC, myData$BIRTHWEIGHT)
```

We can see from the plot that there doesn't seem to be a correlation between the mother's education and the baby's weight at birth. The correlation between the data is $\widehat{\rho_0} = 0.09$.

The null hypothesis is $H_0 : \rho = 0$, there is no correlation between the mother's education level and the baby's weight at birth. The alternate hypothesis is the two-sided $H_1 : \rho \neq 0$, there is a correlation between the data.

The plot of the data did not give us any additional prior knowledge whether there is a correlation or not, hence the two-sided alternate hypothesis. $\alpha = 0.05$.

Here, we will simulate the behavior of $\widehat{\rho}$ when the null hypothesis is true, and then to compare the original $\widehat{\rho}$ to the simulated results.

b

Here, we scramble the column with educational level, and re-compute the statistic $\widehat{\rho}$ for the permuted data set. We are forcing the null hypothesis $H_0 : \rho = 0$ to be true.

```
myData[1] <- myData[1][sample(1:nrow(myData)),]

rho_hat <- cor(myData$YEARSEDUC, myData$BIRTHWEIGHT)
```

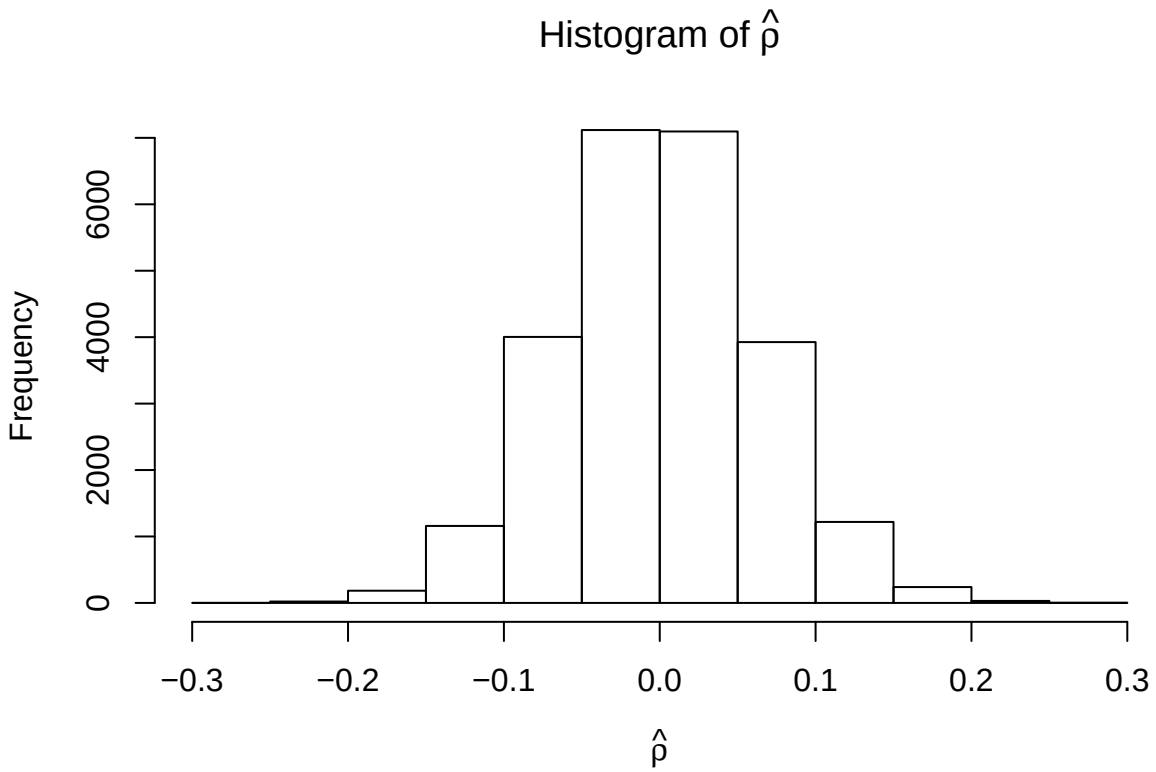
We now have $\widehat{\rho} = -0.032$.

c

We now repat this permutation process 25,000 times and make a histogram of the resulting $\widehat{\rho}$.

```
rhos <- c()
for(i in 1:25000){
  myData[1] <- myData[1][sample(1:nrow(myData)),]
  rhos[i] <- cor(myData$YEARSEDUC, myData$BIRTHWEIGHT)
}

hist(rhos, main=expression(paste("Histogram of ", hat(rho))),
     xlab=expression(hat(rho)))
```

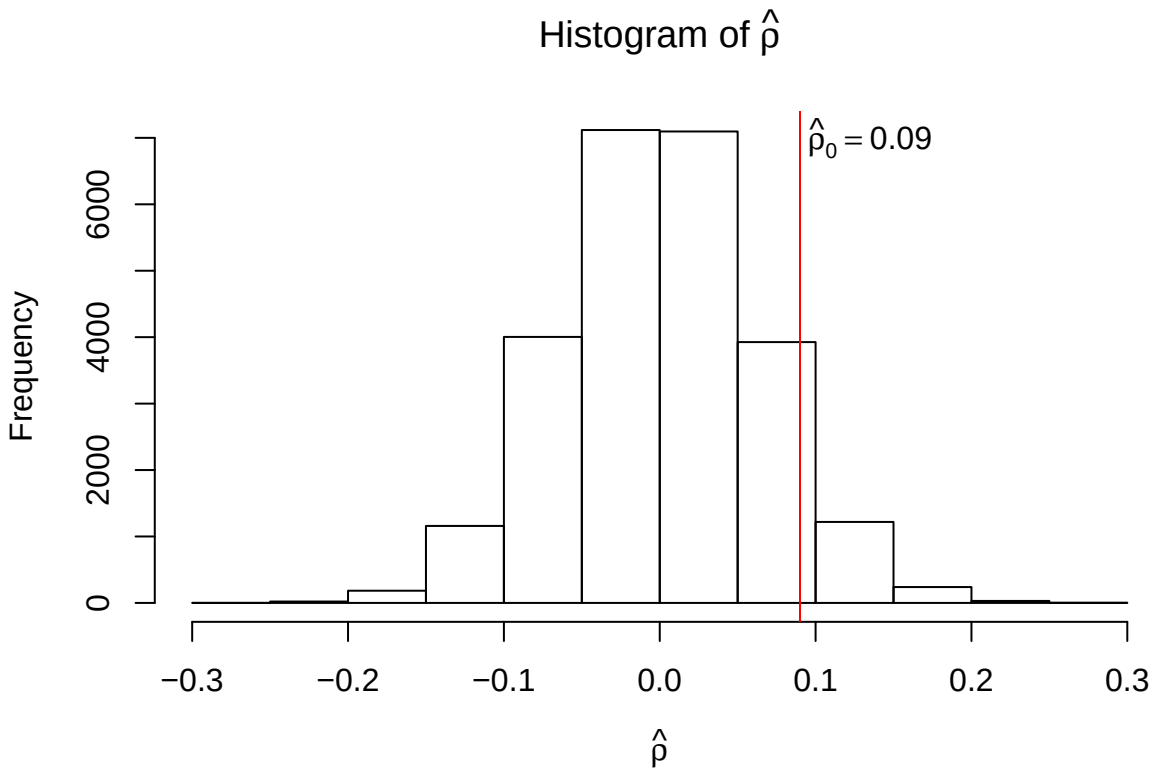


The resulting histogram is centered around 0 and looks somewhat normally distributed. This distribution is called the **randomization distribution** under H_0.

d

We draw the original $\widehat{\rho}$ as a vertical line on the histogram.

```
hist(rhos, main=expression(paste("Histogram of ", hat(rho))),
     xlab=expression(hat(rho)))
abline(v = rho_hat_0, col="red")
text(rho_hat_0*1.5, 7000, bquote(hat(rho)[0] == .(round(rho_hat_0, 3))))
```




```
p_empirical <- sum(rhos > rho_hat_0)/length(rhos)
```

Here, we get $p_{empirical} = 0.08$.

Since we get $2p_{empirical} = 0.16 > \alpha = 0.05$, we do not have enough data at this point in time to reject the null hypothesis that there is no correlation between the data. With a type I error of 5%, $\rho = 0$ and there is no correlation between the mother's education level and the baby's weight at birth.
