Chapter 9: Inferences Based on Two Samples

Curtis Miller

2018-06-29

Introduction

Researchers' questions often address not just one population but two. Frequently the researcher's question doesn't specify a value for a single parameter but gives a relationship between two parameters from two groups so that a relationship can be inferred. For example, while the effect of a new blood pressure drug on blood pressure is good to know, a more interesting question may ask whether a new drug reduces blood pressure more than existing methods.

This chapter discusses procedures intended to compare two samples from two different populations. We will see both how to conduct statistical tests and confidence intervals. The framework for confidence intervals and hypothesis testing hasn't changed. That means that for this chapter a few formulas appropriate for certain contexts are all we need; we don't need to reintroduce the theory.

Section 1: z Tests and Confidence Intervals for a Difference Between Two Population Means

Throughout this chapter I will assume that, unless otherwise stated, we have two different samples, $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ drawn from two independent[1] samples, and that the data within a sample is i.i.d. Let $\mathbb{E}\left[X_1\right] = \mu_X$, $\mathbb{E}\left[Y_1\right] = \mu_Y$, $\text{Var}\left(X_1\right) = \sigma_X^2$, and $\text{Var}\left(Y_1\right) = \sigma_Y^2$.

We're often interested in $\Delta = \mu_X - \mu_Y$. What is an estimator for Δ? Is it unbiased?

What is the variance of this estimator?

[1] In Section 3, the samples are not independent, and $m = n$.

Based on this we can find the sampling distribution for $\hat{\Delta}$ that is at least approximately correct when sample sizes are large.

Assume that σ_X^2 and σ_Y^2 are known. Using the above distribution of $\hat{\Delta}$ we can obtain a confidence interval for the true Δ that is appropriate at least approximately for large m and n.[2]

When planning a study, if we decide in advance to set $m = n$, we could obtain a formula for n (and thus m as well) that will guarantee a chosen margin of error.

[2] Here we will say that m and n are "large" when both quantities are greater than 40.

For hypothesis testing we want to make a statement about the value of Δ. The ingredients of this statistical test are listed below.[3]

Type II error analysis formulas are provided below.

[3] While the test described below is appropriate for any proposed difference Δ_0, the case $\Delta_0 = 0$ is certainly the most interesting and more frequently seen, as this corresponds to the null hypothesis $H_0 : \mu_X = \mu_Y$; in other words the test determines whether the means of the two populations are the same or differ in some way.

If we require that both samples have a common sample size we can also obtain a formula that gives sample sizes that produce a test with a specified Type II error rate for a particular Δ' for a test of significance level α.

Example 1

A tutoring service claims to help understand difficult statistics concepts. You decide to test this. You randomly assign 100 students taking a statistics class to sign up for the tutoring students, while the rest only attend lectures and office hours while learning statistics. At first an equal number of students were assigned to both groups, but after students dropped out, there were 45 students who did not use the tutoring service and 51 who did.[4]

At the end of the course the final exam scores of students from both groups were compared. The students who got tutoring (denoted $X_1, \ldots, X_{51}$) had an average score of 78.79 points. For those who did not get tutoring (denoted $Y_1, \ldots, Y_{45}$), the mean score was 71.09. Assume that $\sigma_X = \sigma_Y = 15$.

1. Compute a 95% confidence interval for the mean difference in scores. Based on this confidence interval, is there good evidence that the tutoring service improves students' performance on exams?

[4] This is known as **dropout bias**; if the propensity to drop out does not depend on whether someone belongs to the control or treatment group, there is no problem, but if there is a relationship the results of a study could be biased. This should be accounted for, but we will ignore the problem.

2. The tutoring service wants the margin of error produced by your study to not exceed 3 points for the exam; this will help the service determine if their product improves students' performance by a letter grade. What sample sizes could achieve this margin of error (while preserving the confidence level)?

3. Perform a statistical test to determine if the tutoring service improves students' scores on exams.

4. The tutoring service wants a statistical test that detects a difference of three points with a Type II error rate of 10% for a test with a Type I error rate of $\alpha = 0.05$. Assuming equal sample sizes for both samples, find sample sizes that leads to a test meeting these requirements.

5. Suppose a statistical has sample sizes of $m = n = 450$, find the Type II error rate when the true difference between the two groups of students is one point.

```
xbar <- 78.79
ybar <- 71.09

sigma_X <- 15
sigma_Y <- sigma_X

m <- 51
n <- 45

alpha <- .05
(z <- qnorm(alpha/2, lower.tail = FALSE))

## [1] 1.959964

# Part 1
(se <- sqrt(sigma_X^2/m + sigma_Y^2/n))

## [1] 3.06786

(moe <- z * se)

## [1] 6.012895

(est <- xbar - ybar)

## [1] 7.7

c(est - moe, est + moe)

## [1]  1.687105 13.712895

# Part 2
ceiling(z^2 * (sigma_X^2 + sigma_Y^2) / 3^2)

## [1] 193

# Part 3
(z_stat <- (est - 0)/se)

## [1] 2.509893

pnorm(z_stat, lower.tail = FALSE)  # p-value

## [1] 0.006038389

# Part 4
ceiling(((sigma_X^2 + sigma_Y^2) *
    (qnorm(.05, lower.tail = FALSE) + qnorm(.1, lower.tail = FALSE))^2/3^2))

## [1] 429
```

```
# Part 5
pnorm(qnorm(.05, lower.tail = FALSE) - 1/
        (sqrt(15^2/450 + 15^2/450)))
```

```
## [1] 0.740489
```

Of course we very rarely know what σ_X and σ_Y are and often need to estimate them from the sample. Then we have an estimated standard error

We would use this for our confidence intervals:

In hypothesis testing our test statistic would be

All the rest is the same. Procedures using these CIs and statistics are appropriate for large sample sizes.

Example 2

The sample standard deviation for the students who got tutoring was 14.52 points. The sample standard deviation for the students who did not get tutoring was 11.87 points. Recompute the confidence interval, test statistic, and p_{val} computed in Example 1.

```
(se_est <- sqrt(14.52^2/m + 11.87^2/n))

## [1] 2.695361

c(est - z * se_est, est + z * se_est)

## [1]  2.417189 12.982811

(z_stat2 <- est/se_est)

## [1] 2.85676

pnorm(z_stat2, lower.tail = FALSE)

## [1] 0.002139946
```

In what contexts can we claim we observe a causal effect in a study? This depends on how the data was generated. If the data was obtained as-is, without being assigned to their groups by the investigators, we may call the study **observational**. If we assigned individuals to groups after the individuals generated their data, we may call the study a **retrospective** observational study. On the other hand, if the investigator assigned individuals randomly to two groups and applied a different treatment depending on group assignment, measuring outcomes after the treatment was applied, we would call the study a **randomized controlled experiment**. The latter type of study allows us to make conclusions about causality, unlike the former.

Section 2: The Two-Sample t Test and Confidence Interval

The procedures from the previous section are appropriate for large sample sizes. When we don't have large sample sizes and we assume the data was drawn from Normal distributions, we can use *t* procedures.

Suppose we assume $\sigma_X = \sigma_Y$. In most cases this assumption is unrealistic, though there are contexts where the assumption makes sense; for instance, we may be attempting to determine not just the difference in mean but whether two samples come from the same population (and thus would have the same population standard deviation). Then the standard error of $\bar{X} - \bar{Y}$ would be

This is estimated with

Consider now the random variable

This random variable follows a t distribution with $m + n - 2$ degrees of freedom. Knowing this, we can find a confidence interval based on this random variable. Procedures that assume that the two samples have the same standard deviation are known as pooled t procedures.

We could also perform a statistical test.

Example 3

Below are two datasets, with each dataset coming from some distribution. Did the same distribution generate these datasets?

```
x <- c( 7.07, -0.01,  8.30,  5.70,  5.06,
        1.85,  0.74,  2.11, -0.93, 15.88)
y <- c( 1.69,  8.83,  9.11, -1.32,  3.97,
        9.40,  7.60,  4.78,  5.13,  6.38)
```

```
mean(x)
```

```
## [1] 4.577
```

```
sd(x)
```

```
## [1] 5.043597
```

```
mean(y)
```

```
## [1] 5.557
```

```
sd(y)
```

```
## [1] 3.472233
```

1. Find a 90% confidence interval for the population mean, using the pooled t procedure.

2. Using the pooled t test, test whether the datasets have the same distribution or not, at significance level $\alpha = 0.1$.

```
t.test(x, y, var.equal = TRUE, conf.level = .9)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = -0.50611, df = 18, p-value = 0.6189
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -4.337743  2.377743
## sample estimates:
## mean of x mean of y
##     4.577     5.557
```

The equal variance assumption made by the pooled test, though, is unrealistic. What if we don't make that assumption? Then we can create procedures based on the quantity

This random variable follows approximately a $t(\nu)$ distribution. The formula for ν is given below:

From this random variable we can derive a CI:

Below is a test statistic for a test based on this random variable:

In most situations you should use the test that does not assume equal variances. When this assumption is true, there is a minor gain in power resulting from using the pooled version of the test, but even a slight deviation from this assumption can result in statistics that behave very poorly; the pooled statistic[5] is not at all robust to the assumption of equal variances.

Example 4

Two teams of competitive rowers locked in a bitter rivalry want to know which team is fastest. Instead of a single race, the two teams each engage in 10 time trials along the same 500m river length. The mean time for team 1 (in minutes) is 1.49 with a standard deviation of 0.12, while for team 2 the mean time was 1.37 with a standard deviation of 0.10.

1. Construct a 95% confidence interval for the difference in rowing times.

2. Perform a statistical test to determine whether team 2 is faster than team 1 or not at significance level $\alpha = 0.1$.

[5] The pooled two-sample test can be viewed as an instance of the likelihood ratio test, while the other test, known as the Wald two-sample t-test, cannot be derived from the likelihood ratio test.

```
x1bar <- 1.49
x2bar <- 1.37

sd1 <- 0.12
sd2 <- 0.10

n <- 10
m <- n

(se1 <- sd1/sqrt(m))

## [1] 0.03794733

(se2 <- sd2/sqrt(n))

## [1] 0.03162278

(std_err_diff <- sqrt(se1^2 + se2^2))

## [1] 0.04939636

(nu <- (se1^2 + se2^2)^2/(se1^4/(m - 1) + se2^4/(n - 1)))

## [1] 17.43311

# Part 1
(tstar <- qt(.975, df = nu))

## [1] 2.10583

c(x1bar - x2bar - tstar * std_err_diff, x1bar - x2bar + tstar * std_err_diff)

## [1] 0.01597968 0.22402032

# Part 2
(test_stat <- (x1bar - x2bar)/std_err_diff)

## [1] 2.429329

pt(test_stat, df = nu, lower.tail = FALSE)

## [1] 0.01309898
```

Section 3: Analysis of Paired Data

Up until now we have required that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two *independent* samples. However, in many experimental designs, the datasets may not be independent; instead, they may be *paired*. That is, we can view the datasets as $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Examples of independent sample studies and paired sample studies are listed below:[6]

We are still primarily interested in $\Delta = \mu_D = \mu_X - \mu_Y$, but since the data is paired, we don't approach inference in the same way. Instead of treating $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ separately, we work with a sample of *differences*:

When we do this, comparing two populations reduces to the one-sample case we saw in chapter 8. Below are CIs and statistical tests in this context:[7]

[6] I list these to avoid a common situation in tests: students confusing paired-sample and independent-sample procedures. Don't be another statistic; *know the difference between these tests!*

[7] What happens when we use the two independent sample procedures in the presence of paired data? The biggest difference is that the variance of our estimator for Δ is no longer correct, since the true variance is

$$\operatorname{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y}{n}$$

For independent samples, $\rho = 0$, but that's likely not the case for paired data; in fact, usually $\rho > 0$. As a result our estimate for the standard error of the statistic is usually larger than appropriate, which makes test statistics smaller than they should be and CIs wider than they should be. While this is "conservative" and thus should be done if you don't know whether data was paired or not, this is usually a major error and needs to be avoided.

Example 5

A drug manufacturer wants to determine if a new weight loss supplement leads to weight loss in subjects. To determine if the supplement leads to weight loss, the manufacturer selects a cohort of six participants to participate. The subjects' weights are measured prior to taking the supplement, then after two months the subjects' weights will be measured again. Below are subjects' weights both before and after taking the supplement:

```
(supp_weight_loss <- data.frame(
  "before" = c(221, 139, 253, 230, 186, 161),
  "after"  = c(209, 121, 230, 220, 182, 162)
))
```

```
##   before after
## 1    221   209
## 2    139   121
## 3    253   230
## 4    230   220
## 5    186   182
## 6    161   162
```

1. Compute the dataset of differences, D_i.

2. Construct a 90% confidence interval for the mean difference in weight loss.

3. Conduct a hypothesis test to determine whether the supplement leads to weight loss. Use a significance level of $\alpha = 0.1$ to decide whether there is a statistically significant difference in weight after taking the supplement.

```
# Compute CI
with(supp_weight_loss,
     t.test(before, after, conf.level = .9, paired = TRUE,
            alternative = "two.sided")
)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 3.0587, df = 5, p-value = 0.02814
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##   3.753293 18.246707
## sample estimates:
## mean of the differences
##                      11
```

```
# Statistical test
with(supp_weight_loss,
     t.test(before, after, paired = TRUE, alternative = "greater")
)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 3.0587, df = 5, p-value = 0.01407
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.753293      Inf
## sample estimates:
## mean of the differences
##                      11
```

Prior to conducting a study, should you opt for an independent-sample study or a paired-sample study? Below are pros and cons of the two:

In general, if there is a lot of variability within a population and a large correlation resulting from pairing, then a paired sample may be preferred to independent samples, while if there is not a lot of variability and the sample is not large, we may prefer an independent-sample procedure.

Section 4: Inferences Concerning a Difference Between Population Proportions

Suppose that instead of being interesterested in the difference between two means of quantitative variables we are interested in the difference in the population proportions from two different populations. For example, we may want to compare the rate at which a disease appears in men versus women, or compare political affiliation across different demographic groups. In any case, there is one population at which the probability of a "success" is p_X and another population where the probability of a "success" is p_Y. If we were conducting a hypothesis test, we may want to see if $p_X = p_Y$ or not.

We say that we have two independent samples of i.i.d. binomial data, $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$. We're interested in estimating $p_X - p_Y$. The natural estimator for this parameter is:

The variance and standard error of this estimator is

When m and n are large, the following random variable follows an approximately standard Normal distribution:

We can use this to construct confidence intervals and statistical tests.

Let's discuss statistical testing first. When H_0 is true, $p_X = p_Y = p$. We need to estimate p, and can do so using the pooled sample (the sample that includes both datasets, $X_1, \dots, X_m, Y_1, \dots, Y_n$). The resulting estimator is

We then have the following test, appropriate for large sample sizes:

Type II errors don't depend on $p_X - p_Y$ but instead on each specific p_X and p_Y, so the Type II error function is denoted with $\beta(p_X, p_Y)$.

The following formulas can be used for Type II error analysis:

When conducting sample size planning, if we set $m = n$ and want to detect a difference in proportions of $p_X - p_Y = d$, after guessing p_X and p_Y we get a guessed sample size of

This is the appropriate sample size for a one-sided alternative hypothesis; for a two-sided alternative, replace α with $\alpha/2$.

Example 6

A pharmaceutical company plans to release a new vaccine intended to reduce the risk of contracting the influenza virus. The company plans to test the vaccine by randomly assigning study participants to a control group and a treatment group. Individuals in the treatment group will receive the new vaccine, while individuals in the control group will receive no treatment.[8] The experiment is conducted in a double-blind fashion; that is, neither patients nor experimental staff will know which patient received which vaccine until after the experiment is complete.

1. The experimenters plan on assigning an equal number of subjects to both control and treatment groups. They want to be able to detect a 5% difference in contraction rate (in the new vaccine's favor) 95% of the time. The current influenza contraction rate is believed to be 20%. The planned significance level is $\alpha = 0.1$. Based on this, what sample size should be used?

2. Using the answer from above, what is the probability of making a

[8] This is ethically suspect, but ignore ethics for now.

Type II error when the new vaccine reduces the rate of influenza contraction by only 1%?

3. When the experiment was actually conducted, after participants left the study for various reasons, the number of individuals in the control group was 886 and the number of individuals in the treatment group was 890. 183 individuals in the control group contracted the flu, while 175 contracted the virus in the treatment group. Test whether the vaccine reduced the occurance of influenza. What is the conclusion?

```
p_X <- .2
d <- .05
alpha <- .1
beta <- .05

# Part 1
(n <- ceiling((qnorm(alpha, lower.tail = FALSE) *
                sqrt(p_X + (p_X - d) * ((1 - p_X) + (1 - (p_X - d)))/2) +
                qnorm(beta, lower.tail = FALSE) * sqrt(p_X * (1 - p_X) +
                                                         (1 - p_X) *
                                                         (1 - (p_X - d)
                                                          )))/d^2))
```

```
## [1] 895
```

```
m <- n

# Part 2
(sigma <- sqrt(p_X * (1 - p_X)/m + (p_X - d) * (1 - (p_X - d))/n))
```

```
## [1] 0.01792286
```

```
(pbar <- (m * p_X + n * (p_X - d))/(m + n))
```

```
## [1] 0.175
```

```
(qbar <- (m * (1 - p_X) + n * (1 - (p_X - d)))/(m + n))
```

```
## [1] 0.825
```

```
pnorm((qnorm(alpha, lower.tail = FALSE) * sqrt(pbar * qbar * (1/m + 1/n)) - d)/
        sigma)
```

```
## [1] 0.06611087
```

```
# Part 3
(phat <- (183 + 175)/(886 + 890))
```

```
## [1] 0.2015766
```

```
(z <- (183/886 - 175/890)/sqrt(phat * (1 - phat) * (1/886 + 1/890)))
```

```
## [1] 0.5208789
```

```
pnorm(z, lower.tail = FALSE)  # p-value
```

```
## [1] 0.3012256
```

We can construct a large-sample confidence interval for the difference in proportions using the formula:[9]

Example 7

Construct a 90% CI for the difference in influenza contraction rates based on the data in Example 6. Does this CI agree with the conclusion of the test?[10]

[9] This interval should be appropriate when $m\hat{p}_X$, $n\hat{p}_Y$, $m(1-\hat{p}_X)$, and $n(1-\hat{p}_X)$ are all at least 10.

[10] Looking at the formulas for the test statistic and the CI, we should not believe that the CI will necessarily agree with the test.

```
(phat_X <- 183/886)
```

```
## [1] 0.2065463
```

```
(phat_Y <- 175/890)
```

```
## [1] 0.1966292
```

```
m <- 886
n <- 890

(se <- qnorm(.05, lower.tail = FALSE) * sqrt(phat_X * (1 - phat_X)/m +
                                              phat_Y * (1 - phat_Y)/n))
```

```
## [1] 0.03131543
```

```
c("Lower" = phat_X - phat_Y - se, "Upper" = phat_X - phat_Y + se)
```

```
##       Lower       Upper
## -0.02139837  0.04123249
```

Section 5: Inferences Concerning Two Population Variances

So far we have been interested in comparing μ_X and μ_Y or p_X and p_Y. Sometimes, though, we may be interested in comparing σ_X^2 and σ_Y^2.

Let $N \sim \chi^2(\nu_n)$ and $D \sim \chi^2(\nu_d)$. Consider the random variable

This random variable follows the $F(\nu_n, \nu_d)$ distribution. The density curve for this distribution is illustrated below:

The pdf and cdf of the $F(\nu_n, \nu_d)$ distribution is difficult to describe but I list the expected value and variance of this distribution below:

Suppose $F \sim F(\nu_n, \nu_d)$. Let f_{α,ν_n,ν_d} satisfy $\mathbb{P}\left(F \geq f_{\alpha,\nu_n,\nu_d}\right) = \alpha$. We will call f_{α,ν_n,ν_d} a critical value of the $F(\nu_n, \nu_d)$ distribution. Critical values (and thus some values of the cdf of) the $F(\nu_n, \nu_d)$ distribution are listed in Table A.9, for select ν_n and ν_d. ν_n is called the **numerator degrees of freedom** and ν_d is called the **denominator degrees of freedom**. An important identity for critical values of the $F(\nu_n, \nu_d)$ distribution is

Example 8

Let $F \sim F(4, 9)$.

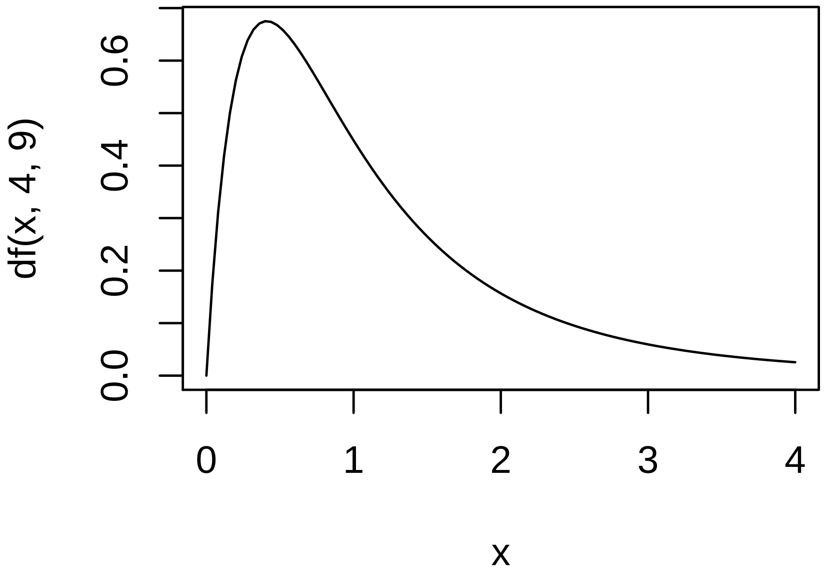
1. Compute $\mathbb{E}[F]$ and $\text{Var}(F)$.

2. Compute $\mathbb{P}(F \leq 3.63)$.

3. Find $f_{.01,4,9}$.

4. Find $f_{.999,4,9}$

```
curve(df(x, 4, 9), 0, 4)
```



```
# Part 1
(mu_f <- integrate(function(x) {x * df(x, 4, 9)}, 0, Inf))

## 1.285714 with absolute error < 7.8e-06

(var_f <- integrate(function(x) {(x - mu_f$value)^2 * df(x, 4, 9)}, 0, Inf))

## 1.818367 with absolute error < 5.5e-05

# Part 2
pf(3.63, 4, 9)

## [1] 0.9498937

# Part 3
qf(.01, 4, 9, lower.tail = FALSE)

## [1] 6.422085

# Part 4
qf(.999, 4, 9, lower.tail = FALSE)

## [1] 0.0206294
```

The F distribution matters because when $X_1, \ldots, X_m$ is an i.i.d. sample with $X_1 \sim N(\mu_X, \sigma_X)$ and $Y_1, \ldots, Y_n$ is an i.i.d. sample with $Y_1 \sim N(\mu_Y, \sigma_Y)$, if S_X^2 is the sample variance for the first dataset and S_Y^2 is the sample variance for the second dataset, we can find a distribution for S_X^2 / S_Y^2.

This distribution can be used for deriving confidence intervals and statistical tests for σ_X^2/σ_Y^2 and thus make statements about the relationship between σ_X^2 and σ_Y^2.[11]

Below I describe a hypothesis test for checking the relationship between σ_X^2 and σ_Y^2.

We can also derive formulas for the confidence interval for σ_X^2/σ_Y^2.

[11] Thus we also have statements for σ_X and σ_Y's relationship when we take square roots appropriately.

Example 9

The standard deviation of the returns of a stock is called the stock's volatility in finance. Two stocks, CGM and UOU, are believed to have Normally distributed returns. Some returns of these stocks are listed below:

```
cgm <- c(-0.004,  0.006,  0.002, -0.023, -0.006, -0.004,  0.023, -0.011,
          0.001,  0    , -0.004)
uou <- c( 0,     -0.011,  0.015,  0.005, -0.012,  0.003, -0.009,  0.005)
```

```
format(var(cgm), scientific = FALSE)
```

```
## [1] "0.0001267636"
```

```
format(var(uou), scientific = FALSE)
```

```
## [1] "0.00008971429"
```

1. Find a 90% confidence interval for $\sigma_{\text{CGM}}/\sigma_{\text{UOU}}$. Based on this CI, is it plausible that the two stocks have the same volatility?

2. Perform a statistical test to check whether the two stocks have the same volatility. Does the result of the test agree with the confidence interval's conclusion?

```
(res <- var.test(cgm, uou, conf.level = .9))

##
##  F test to compare two variances
##
## data:  cgm and uou
## F = 1.413, num df = 10, denom df = 7,
## p-value = 0.6643
## alternative hypothesis: true ratio of variances is not equal to 1
## 90 percent confidence interval:
##  0.3885498 4.4303192
## sample estimates:
## ratio of variances
##            1.41297

sqrt(res$conf.int)  # Need to take square root to get CI of volatility ratio

## [1] 0.6233377 2.1048323
## attr(,"conf.level")
## [1] 0.9
```