**Inference for two variances**



We previously compared two means through subtraction. This section compares two variances through a ratio. A ratio is used here because the probability distribution for corresponding statistics is easier to derive.

Examples:

* In comparison of two stocks, which has more risk? Remember that σ2 is used as a measurement of investment risk.
* Suppose two types of cars claim the same miles per gallon of gas. Is one car type more consistent?
* To compare two frozen pizza manufacturing facilities, random samples of pizza weights are taken. Are the facilities producing pizzas that have the same consistency in their weights?

The ratio of two population variances is , where  is the variance for population #1 and  is the variance for population #2.

Notes:

* If  = 1, then  = 
* If  > 1, then  > ; there is more variability in population #1 than population #2.
* If  < 1, then  < ; there is less variability in population #1 than population #2.

F probability distribution function

In order to perform inference with respect to , we need to use the F probability distribution. This is another probability distribution which is often used in statistics. Below is its definition:

The continuous random variable X has a F probability distribution, with ν1 and ν2 degrees of freedom (two separate values), if its mathematical function is given by



where ν1 and ν2 are greater than 0.

The values of ν1 and ν2 control the shape of the distribution. It will be clear shortly why there are two different degrees of freedom. One of these degrees of freedom will correspond to a statistic in the numerator (ν1) of a fraction and one will correspond to the denominator (ν2) of a fraction.

Mean and variance of a random variable with this distribution:

E(X) =  for ν2 > 2 and

Var(X) =  for ν2 > 4

Notation:  denotes the 1 – α/2 quantile from a F distribution with ν1 and ν2 degrees of freedom.



Then P(< X < ) = 1 – α. Note that  = , which will be used shortly.

Example: F probability distribution plot (F\_distribution.xlsx)

This is an interactive file which allows you to see the probability distribution for different degrees of freedom.



Example: Finding probabilities and quantiles from a F distribution (F\_dist.R)

To find P(X < 2.71) with ν1 = 5 and ν2 = 20, we could use integration:



Instead, we will use the pf() function:

> pf(q = 2.71089, df1 = 5, df2 = 20)

[1] 0.95

To find the 1 – α/2 quantile from a F distribution, we could use integration:



where we would solve for c in the above equation. Instead, we will use the qf() function:

> alpha <- 0.1

> qf(p = 1 - alpha/2, df1 = 5, df2 = 20)

[1] 2.71089

The df() function allows us to evaluate f(x) so that we can plot the distribution:

> curve(expr = df(x = x, df1 = 5, df2 = 20), xlim = c(0,7),

 col = "red", lwd = 2, main = "F distribution with

 nu1 = 5 and nu2 = 20 DF", ylab = "f(x)", xlab = "x", n

 = 1000)

> abline(h = 0)



Some introduction to statistics books provide probabilities corresponding to particular degrees of freedoms in a table format. We will not use a table in this course.

Probability distribution involving 

The natural way to estimate  is with , so we would like to have a probability distribution involving this statistic. In order to do this, we need to make the assumption that random samples of size nj from both populations are characterized by normal probability distributions with E(Yj) = μj and Var(Yj) =  for population j = 1 and 2. With this assumption, one can show that



has an F probability distribution with ν1 = n1 – 1 and ν2 = n2 – 1 degrees of freedom.

The reason why a probability distribution for  is of interest to us is because it allows us to make statements such as



Rearranging some terms in the probability expression gives us









This leads to a confidence interval for .

CI for  – If  and  are the observed variances of a random sample of size n1 and n2, respectively, from two populations characterized by normal distributions, a (1 –α)100% CI for  is



Notes:

* Taking the positive square root leads to a confidence interval for .
* We could also use the formula:



to obtain the same interval.

Example: ConAgra (ConAgra.R, ConAgra\_SP500.csv)

Suppose an investor has a choice between investing in shares of ConAgra, a food company, or an S&P 500 Index mutual fund. The investor prefers the investment with the higher weekly mean return and lower weekly risk.

The sample consists of weekly returns of ConAgra stock and the S&P 500 Index mutual fund (just the actual S&P500) over a 1-year period of time:

The blue highlighted text below is a small correction to the notes.

|  |  |  |
| --- | --- | --- |
| **Week** | **S&P 500** | **ConAgra** |
| 1 | -0.27% | 0.58% |
| 2 | -2.87% | -5.91% |
|  |  |  |
| 51 | -1.82% | 6.51% |
| 52 | 7.32% | 3.80% |

For example, ConAgra shareholders lost 5.91% of their investment during the second week. Suppose an investor had $1,000 worth of ConAgra shares at the end of the first week. By the end of the second week, the investment had declined to $1000×(1 + 0.0058)×(1 – 0.0591) = $946.36.

Is this a random sample? If not, can we still use it?

Although this data is not a random sample from the populations, there is a very low amount of dependence between successive data values. One way to show this is to perform a “Durbin-Watson Test”. This test showed there is little dependence among the data.

Plots:

> conagra <- read.csv(file = "ConAgra\_SP500.csv")

> head(conagra)

 week sp500 conagra

1 1 -0.0027 0.0058

2 2 -0.0287 -0.0591

3 3 0.0141 0.0749

4 4 0.0009 0.0414

5 5 0.0374 0.0777

6 6 -0.0080 -0.0336

> tail(conagra)

 week sp500 conagra

47 47 0.0361 0.0023

48 48 0.0109 0.0463

49 49 0.0242 -0.0066

50 50 -0.0403 -0.0089

51 51 -0.0182 0.0651

52 52 0.0732 0.0380

> boxplot(x = cbind(conagra$sp500, conagra$conagra), main =

 "Box and dot plot", ylab = "Return", xlab =

 "Investment", pars = list(outpch=NA), names = c("S&P

 500", "ConAgra"), col = NA)

> stripchart(x = conagra$sp500, lwd = 2, col = "red",

 method = "jitter", vertical = TRUE, pch = 1, add =

 TRUE, at = 1)

> stripchart(x = conagra$conagra, lwd = 2, col = "red",

 method = "jitter", vertical = TRUE, pch = 1, add =

 TRUE, at = 2)



Due to the structure of the data, I had to use a little different code than in the past to produce the box and dot plots. Alternatively, one could restructure the data to a similar format as we have seen before and use the usual code for the plots. This alternative is shown in my corresponding program.

> #Specify the breaks to make sure they are the same for

 both plots

> par(mfrow = c(2,1))

> hist(x = conagra$sp500, main = "S&P 500", xlab =

 "Return", breaks = seq(from = -0.15, to = 0.1, by =

 0.05), col = NA)

> hist(x = conagra$conagra, main = "ConAgra", xlab =

 "Return", breaks = seq(from = -0.15, to = 0.1, by =

 0.05), col = NA)



> mean(conagra$sp500)

[1] 0.002495019

> mean(conagra$conagra)

[1] 1.340385e-05

> s1.sq <- var(conagra$sp500)

> s2.sq <- var(conagra$conagra)

> data.frame(s1.sq, s2.sq)

 s1.sq s2.sq

1 0.0006777131 0.002314334

Questions:

* Is there much of a difference in the mean return?
* Do you think the variances are equal? If not, which investment do you think has the larger variance?
* What do you think about normal probability distribution assumptions here?

Let 1 = S&P 500 and 2 = ConAgra.

Below are the calculations for a 95% confidence interval:



> alpha <- 0.05

> qf(p = alpha/2, df1 = n2 - 1, df2 = n1 - 1)

[1] 0.574

> qf(p = 1 - alpha/2, df1 = n2 - 1, df2 = n1 - 1)

[1] 1.742

> n1 <- length(conagra$sp500)

> n2 <- length(conagra$conagra)

> data.frame(n1, n2)

 n1 n2

1 52 52

> lower <- s1.sq/s2.sq \* qf(p = alpha/2, df1 = n2 - 1, df2

 = n1 - 1)

> upper <- s1.sq/s2.sq \* qf(p = 1 - alpha/2, df1 = n2 - 1,

 df2 = n1 - 1)

> data.frame(lower, upper)

 lower upper

1 0.1681 0.5101

Below is an easier way to perform the calculations in R:

> var.test(x = conagra$sp500, y = conagra$conagra,

 conf.level = 0.95)

 F test to compare two variances

data: conagra$sp500 and conagra$conagra

F = 0.2928, num df = 51, denom df = 51, p-value = 2.255e-05

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

 0.1681 0.5101

sample estimates:

ratio of variances

 0.2928

What do the results here say about the risk of the two investments?

What about hypothesis testing?

Tests can be performed using confidence intervals, test statistics, and p-values in a similar manner as we saw in the previous chapters. We just now use different formulas for confidence intervals and test statistics!

Example: ConAgra (ConAgra.R, ConAgra\_SP500.csv)

The investor is looking for one investment to choose. Suppose the investor hypothesizes that the S&P 500 mutual fund would be the less risky investment and wants to determine if there is evidence to confirm his hypothesis. For this situation, we would want to put  then in the alternative hypothesis so that we can control the probability of incorrectly concluding this is true with a stated error level.

The hypotheses are

Ho:  ≥ 1

Ha:  < 1

CI Method using α = 0.05:

1. Ho:  ≥ 1
Ha:  < 1
2. A one-sided CI



Question: Why is “α” in the subscript for F instead of “α/2”?

Below are the calculations from R:

> s1.sq/s2.sq \* qf(p = 1 - alpha, df1 = n2 - 1, df2 = n1

 - 1)

[1] 0.4662

The 95% confidence interval is 0 <  < 0.4662.

1. Reject Ho because 1 is not in the interval.
2. There is sufficient evidence to conclude that the S&P 500 mutual fund is a less risky investment compared to ConAgra.

Of course, you could also perform the hypothesis test using the test statistic and p-value methods.

Test statistic: Earlier, we saw that



Notice that we can re-write the middle portion of the inequality as



This leads then to a test statistic of



if  is contained within the null hypothesis. If instead  for some positive constant c, then



What are the critical values for two-tail, left-tail, and right-tail tests?

P-value for two-tail test:  where X is a random variable that has a F distribution with ν1 = n1 – 1 and ν2 = n2 – 1 degrees of freedom.

Questions:

* Why isn’t the p-value 2?
* What is the p-value formula for a left-tail or right-tail test?

Example: ConAgra (ConAgra.R, ConAgra\_SP500.csv)

Test statistic method using α = 0.05:

1. Ho:  ≥ 1
Ha:  < 1
2. 

> F.stat <- s1.sq/s2.sq

> F.stat

[1] 0.2928

1. 

> qf(p = alpha, df1 = n1 - 1, df2 = n2 - 1)

[1] 0.6282

1. 
Reject Ho because 0.2928 < 0.6282
2. There is sufficient evidence to conclude that the S&P 500 mutual fund is a less risky investment compared to ConAgra.

P-value method using α = 0.05:

1. Ho:  ≥ 1
Ha:  < 1
2. P(X < 0.2928) = 0.0000128

> pf(q = F.stat, df1 = n2 - 1, df2 = n1 - 1)

[1] 1.128e-05

1. α = 0.05
2. Reject Ho because 0.0000128 < 0.05
3. There is sufficient evidence to conclude that the S&P 500 mutual fund is a less risky investment compared to ConAgra.

Below is an easier way to perform the calculations in R:

> var.test(x = conagra$sp500, y = conagra$conagra, ratio =

 1, alternative = "less", conf.level = 0.95)

 F test to compare two variances

data: conagra$sp500 and conagra$conagra

F = 0.2928, num df = 51, denom df = 51, p-value = 1.128e-05

alternative hypothesis: true ratio of variances is less than 1

95 percent confidence interval:

 0.0000 0.4662

sample estimates:

ratio of variances

 0.2928

While the methods presented here for inference on  are the most widely used, the normal distribution assumption needed for the confidence interval and the hypothesis test is a potential problem. There are other ways to find a confidence interval and perform a hypothesis test. One statistical procedure that can be used is the bootstrap.