**Inferences about μ1 - μ2: independent samples**



The purpose of this section is to compare the means of two populations. The samples from the two populations are taken independently of each other.

Examples:

* Who receives higher average grades?
	+ Males or Females?
	+ Athletes or Non-athletes?
	+ Greeks or Non-greeks?
* Which drug helps relieve allergy symptoms better on average: Claritin or Allegra?
* An economist wishes to determine whether there is a difference in mean family income for households in two socioeconomic groups.

Another version of the central limit theorem: If independent samples of size n1 and n2 are drawn at random from two populations with means μ1 and μ2 and variances  and  then the sampling distribution of the difference between sample means, -, is approximately a normal probability distribution with



and

.

Thus,



has an approximate standard normal distribution for large n1 and n2.

Notes:

* n1≥30 and n2≥30 is usually a large enough sample so that the central limit theorem holds.
* Probabilities can be found using the normal probability distribution here in a similar manner as done when there was only one sample mean.

How can we use this result to find a CI for μ1-μ2?

The same way as we did for one population mean! Note that



After rearranging terms, we find



CI for μ1-μ2 with  and  known: If  and  are the observed means of independent random samples from two populations, respectively, an approximate (1-α)100% CI for μ1-μ2 is given by



where zα/2 is the 1 – α/2 quantile from a standard normal distribution.

Notes:

* Nothing needs to be said about the probability distribution functions for the two populations. The word “approximate” is used because the central limit theorem is needed. What does this mean for the sample sizes required for this to work?
* The (1-α)100% confidence level means that in repeated sampling, (1-α)100% of the formed confidence intervals will actually contain μ1-μ2.
* Problem: How do you know what  and  are? In actual applications, you will not so this CI is actually of very limited use!

There are two things that can be done to get around knowing  and :

1. Assume  =  = σ2 (say) and estimate σ2 with a pooled estimator denoted by .

With normal distribution assumptions for both populations, the t distribution can be used to derive a CI. The interval becomes



where . Still, the  =  = σ2 is bothersome so we will not use this interval.

1. Use  and  in the CI derived through the central limit theorem and replace the standard normal distribution quantiles with t distribution quantiles.

This CI also needs the populations to be characterized by normal distributions similar to 1). If the sample sizes are large, this assumption will not be important. We will use this interval for μ1-μ2.

If we have two populations that can be characterized by a normal distribution, then



has an approximate t distribution with  as the sample variance for sample i from population i=1,2 and degrees of freedom of



Notes:

* The degrees of freedom are derived here through what is called a “Satterthwaite adjustment”. Do not worry about its derivation. You may learn about deriving this adjustment in a mathematical statistics course. Also, you will most likely see it applied again in experimental design course.
* Usually, ν will not be an integer. Some books will say to truncate (round down) a degrees of freedom that is not an integer. This is because the authors only use t distribution tables. R has no problems with non-integer degrees of freedom, so we will not truncate.
* One can show that



and this will lead to a CI

CI for μ1-μ2 with  and  unknown and possibly unequal: If ,  and ,  are the observed means and variances of independent random samples of size n1 and n2, respectively, from populations with normal probability distributions and possibly unequal and unknown variances, then an approximate (1-α)100% CI for μ1-μ2 is



where  is the 1 – α/2 quantile from a t distribution where with



degrees of freedom.

Example: Dividend Yield (div\_yield\_inference.R; div\_yield.csv)

Is there a difference in average dividend yield of companies traded on the NASDAQ vs. NYSE? Find a 95% CI to estimate the mean difference.

Below is part of the data set:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number | Stock | **Stock Exchange** | **Closing Price** | **Dividend** | **Dividend Yield** |
| 1 | AMF Bowlng | NYSE | 21.75 | 0 | 0.00% |
| 2 | Alr TOPRS | NYSE | 25.25 | 1.9 | 7.52% |
| 3 | AmerHess | NYSE | 62.1875 | 0.6 | 0.96% |
|  |  |  |  |  |  |
| 30 | HighwdProp pfB | NYSE | 24.5625 | 0 | 0.00% |

| Number | Stock | **Stock Exchange** | **Closing Price** | **Dividend** | **Dividend Yield** |
| --- | --- | --- | --- | --- | --- |
| 1 | AccuMed | NASDAQ | 2.25 | 0 | 0.00% |
| 2 | AG Assoc | NASDAQ | 5.75 | 0 | 0.00% |
| 3 | AllsFnl | NASDAQ | 5.625 | 0 | 0.00% |
|  |  |  |  |  |  |
| 30 | GalileoCp | NASDAQ | 11.375 | 0 | 0.00% |

Below are the plots and numerical summaries:





> div <- read.csv(file = "div\_yield.csv")

> head(div) #Shows first 6 observations

 ID Company Exchange Closing\_Price Dividend Dividend\_Yield

1 1 AMF Bowlng NYSE 21.75 0.00 0.000000

2 2 Alr TOPRS NYSE 25.25 1.90 0.075248

3 3 AmerHess NYSE 62.19 0.60 0.009648

4 4 AmStratll NYSE 11.81 0.99 0.083810

5 5 ArdenRlty NYSE 30.75 1.60 0.052033

6 6 Aviall NYSE 13.31 0.00 0.000000

> aggregate(x = Dividend\_Yield ~ Exchange, data = div,

 FUN = mean)

Added after video recording: R has changed the syntax for aggregate(). In the video, I show formula = Dividend\_Yield ~ Exchange. Now, the proper syntax is x = Dividend\_Yield ~ Exchange. I made the correction here and in the program.

 Exchange Dividend\_Yield

1 NASDAQ 0.01180

2 NYSE 0.02060

> aggregate(x = Dividend\_Yield ~ Exchange, data = div,

 FUN = sd)

 Exchange Dividend\_Yield

1 NASDAQ 0.02894

2 NYSE 0.03191

> aggregate(x = Dividend\_Yield ~ Exchange, data = div,

 FUN = quantile, probs = c(0.25, 0.5, 0.75))

 Exchange Dividend\_Yield.25% Dividend\_Yield.50% Dividend\_Yield.75%

1 NASDAQ 0.00000 0.00000 0.01282

2 NYSE 0.00000 0.00000 0.04308

Questions:

* Do you think there is any preliminary evidence of a difference between the mean dividend yields for all companies traded on the stock exchanges?
* What do you think about the normal probability distribution approximation for each population?
* Is the normal probability distribution approximation for each population needed?

Below are some calculations from R:

> alpha <- 0.05

> #Let 1 = NASDAQ and 2 = NYSE

> save.mean <- aggregate(x = Dividend\_Yield ~

 Exchange, data = div, FUN = mean)

> ybar1 <- save.mean$Dividend\_Yield[1]

> ybar2 <- save.mean$Dividend\_Yield[2]

> save.var <- aggregate(x = Dividend\_Yield ~

 Exchange, data = div, FUN = var)

> s.sq1 <- save.var$Dividend\_Yield[1]

> s.sq2 <- save.var$Dividend\_Yield[2]

> save.n <- aggregate(x = Dividend\_Yield ~

 Exchange, data = div, FUN = length)

> n1 <- save.n$Dividend\_Yield[1]

> n2 <- save.n$Dividend\_Yield[2]

> #################

> #Variance unequal

> nu <- (s.sq1/n1 + s.sq2/n2)^2 / ( (s.sq1/n1)^2 /

 (n1 - 1) + (s.sq2/n2)^2 / (n2 - 1))

> data.frame(ybar1, ybar2, s.sq1, s.sq2, n1, n2, nu,

 t.quant = qt(p = 1 - alpha/2, df = nu))

 ybar1 ybar2 s.sq1 s.sq2 n1 n2 nu t.quant

1 0.0118 0.0206 0.0008374 0.001018 30 30 57.45 2.002

> lower <- ybar1 - ybar2 - qt(p = 1 - alpha/2, df = nu) \*

 sqrt(s.sq1/n1 + s.sq2/n2)

> upper <- ybar1 - ybar2 + qt(p = 1 - alpha/2, df = nu) \*

 sqrt(s.sq1/n1 + s.sq2/n2)

> data.frame(lower, upper)

 lower upper

1 -0.02454 0.006952

Suppose 1 = NASDAQ and 2 = NYSE. Thus, the approximate 95% CI for μ1-μ2 is:





-0.0245< μ1 - μ2 < 0.0070

I am 95% confident that the difference in dividend yields between the NASDAQ and NYSE is between -2.45% and 0.7%.

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

> t.test(formula = Dividend\_Yield ~ Exchange, data = div,

 var.equal = FALSE, conf.level = 0.95)

 Welch Two Sample t-test

data: Dividend\_Yield by Exchange

t = -1.118, df = 57.45, p-value = 0.2682

alternative hypothesis: true difference in means is not

 equal to 0

95 percent confidence interval:

 -0.024539 0.006952

sample estimates:

mean in group NASDAQ mean in group NYSE

 0.0118 0.0206

Notice that NASDAQ’s mean is listed first at the end of the output. This means that the first population mean μ1 represents NASDAQ.

Questions:

* Is there sufficient evidence to indicate a difference in the mean dividend yields between companies listed on the two stock exchanges?
* What are the populations?
* What would you expect the plots to look like if there was enough evidence to indicate a difference between the means?

Remember that the confidence level DOES NOT MEAN there is a 95% probability that this CI contains μ1-μ2. Instead, it means if many more similar samples were taken and corresponding confidence intervals calculated, we would expect on average about 95% of the CIs to contain μ1 - μ2.

Suppose the data was in the following form:

> nasdaq <- div$Dividend\_Yield[div$Exchange == "NASDAQ"]

> nyse <- div$Dividend\_Yield[div$Exchange == "NYSE"]

> head(nasdaq)

[1] 0.000000 0.000000 0.000000 0.151450 0.011034 0.006115

> head(nyse)

[1] 0.000000 0.075248 0.009648 0.083810 0.052033 0.000000

The t.test() function can still be used to perform the test. There are now x and y arguments where one can specify the data:

> t.test(x = nasdaq, y = nyse, var.equal = FALSE,

 conf.level = 0.95)

 Welch Two Sample t-test

data: nasdaq and nyse

t = -1.118, df = 57.45, p-value = 0.2682

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -0.024539 0.006952

sample estimates:

mean of x mean of y

 0.0118 0.0206

What about hypothesis testing?

Tests can be performed using confidence intervals, test statistics, and p-values in a similar manner as we saw for one population mean. We just now use different formulas for confidence intervals and test statistics than we did in the last chapter.

Example: Dividend Yield (div\_yield\_inference.R)

Is there a difference in average dividend yield of companies traded on the NYSE vs. NASDAQ? Perform a hypothesis test to determine if there is a difference.

Suppose 1 = NASDAQ and 2 = NYSE.

CI Method using α = 0.05:

* 1. Ho:μ1 - μ2 = 0
	Ha:μ1 - μ2 ≠ 0
	2. The 95% CI is -0.0245< μ1 - μ2 < 0.0070
	3. Do not reject Ho since 0 is in the interval.
	4. There is not sufficient evidence to indicate a difference between the mean dividend yields for companies traded on the two stock exchanges.

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

Test statistic: Earlier, we saw that



where 

Notice that μ1 - μ2 is almost always replaced with 0 to reflect a null hypothesis test of Ho:μ1 - μ2 = 0 or Ho:μ1 - μ2 ≤ 0 or Ho:μ1 - μ2 ≥ 0.

What is the test statistic?

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

P-value for two-tail test: 2×P(T > |t|) where T has ν degrees of freedom.

What is the p-value for a left-tail or right-tail test?

Example: Dividend Yield (div\_yield\_inference.R)

R code and output:

> qt(p = 1 - alpha/2, df = nu)

[1] 2.002

> t.test(formula = Dividend\_Yield ~ Exchange, data =

 div, var.equal = FALSE, conf.level = 0.95, mu = 0)

 Welch Two Sample t-test

data: Dividend\_Yield by Exchange

t = -1.118, df = 57.45, p-value = 0.2682

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -0.024539 0.006952

sample estimates:

mean in group NASDAQ mean in group NYSE

 0.0118 0.0206

For emphasis, I included the mu = 0 default argument value in the code. This corresponds to μ1 - μ2 = 0 in the null hypothesis.

Test statistic method (using α = 0.05):

1. Ho:μ1 - μ2 = 0
Ha:μ1 - μ2 ≠ 0
2. 
3. ±t0.025,57.45 = ±2.002
4. 

Do not reject Ho because -2.002 < -1.1181 < 2.002
5. There is not sufficient evidence to indicate a difference between the mean dividend yields for companies traded on the two stock exchanges.

P-value method:

1. Ho:μ1 - μ2 = 0
Ha:μ1 - μ2 ≠ 0
2. 2×P(T > |-1.1181|) = 0.2682
3. α = 0.05
4. Do not reject Ho because 0.2682 > 0.05
5. There is not sufficient evidence to indicate a difference between the dividend yields for companies traded on the two stock exchanges.

Notes:

* If desired, the results from t.test() can be saved into a function so that parts of the results can be used separately:

> save.test <- t.test(formula = Dividend\_Yield ~

 Exchange, data = div, var.equal = FALSE, conf.level =

 0.95, mu = 0)

> names(save.test)

[1] "statistic" "parameter" "p.value" "conf.int"

 "estimate"

[6] "null.value" "alternative" "method" "data.name"

> save.test$statistic

 t

-1.118

* If a left-tail test of Ho:μ1 - μ2 ≥ 0 vs. Ha:μ1 - μ2 < 0 was performed:

> t.test(formula = Dividend\_Yield ~ Exchange, data = div,

 alternative = "less", var.equal = FALSE, conf.level =

 0.95, mu = 0)

 Welch Two Sample t-test

data: Dividend\_Yield by Exchange

t = -1.118, df = 57.45, p-value = 0.1341

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

 -Inf 0.004354

sample estimates:

mean in group NASDAQ mean in group NYSE

 0.0118 0.0206