

## Two-Sample t-Procedures for Means

DS705

# Pooled vs. Unpooled

Pooled-variance t-procedures

- ▶ `var.equal=TRUE`

Separate-variance (Welch, Unpooled) t-procedures

- ▶ `var.equal=FALSE`

## Two-Sample t-Procedures for Means

### └ Pooled vs. Unpooled

#### Pooled vs. Unpooled

##### Pooled-variance t-procedures

- `var.equal=TRUE`

##### Separate-variance (Welch, Unpooled) t-procedures

- `var.equal=FALSE`

Pooled vs Unpooled – that is the question when it comes to two-sample t-tests and intervals for population means.

Pooled and unpooled procedures both require independent, random samples from normally distributed populations.

The difference between the two is that the pooled procedures have the additional data requirement that the populations have equal variances. In R, the option to specify this in the `t.test` function is `var.equal=TRUE`.

The separate-variance or unpooled t-procedures require no such stipulation that the population variances be equal. In R, the option to specify this in the `t.test` function is `var.equal=FALSE` and this is the default case if the `var.equal` option is omitted.

## Interpreting CI's for a Difference: $\mu_1 - \mu_2 = ?$

The importance of 0 in the interval . . . or not in the interval

## Two-Sample t-Procedures for Means

## └ Interpreting CI's for a Difference:

 $\mu_1 - \mu_2 = ?$ Interpreting CI's for a Difference:  $\mu_1 - \mu_2 = ?$ 

The importance of 0 in the interval . . . or not in the interval

With the two-sample t-intervals, the purpose is to estimate the difference between two unknown population means. We're estimating the answer to a subtraction problem, and so of course it will be important to determine if the difference is positive, negative, or if it could be 0.

## Interpreting CI's for a Difference: $\mu_1 - \mu_2 = ?$

The importance of 0 in the interval . . . or not in the interval.

Then for a given level of confidence:

If  $(-3, 2)$ , then "... the difference in population means is between -3 and 2..."

## Two-Sample t-Procedures for Means

## └ Interpreting CI's for a Difference:

 $\mu_1 - \mu_2 = ?$ Interpreting CI's for a Difference:  $\mu_1 - \mu_2 = ?$ 

The importance of 0 in the interval . . . or not in the interval.

Then for a given level of confidence:

If  $(-3, 2)$ , then "...the difference in population means is between -3 and 2..."

When 0 is in the interval, which is evident when the lower bound is negative and the upper bound is positive, then there is insufficient evidence to conclude that the population means are different, such as in this example here.

## Interpreting CI's for a Difference: $\mu_1 - \mu_2 = ?$

The importance of 0 in the interval . . . or not in the interval.

Then for a given level of confidence:

If  $(-3, 2)$ , then "...the difference in population means is between -3 and 2..."

If  $(-7, -4)$ , then "...population mean 1 is 4 to 7 units less than population mean 2..."



## Two-Sample t-Procedures for Means

## └ Interpreting CI's for a Difference:

 $\mu_1 - \mu_2 = ?$ Interpreting CI's for a Difference:  $\mu_1 - \mu_2 = ?$ 

The importance of 0 in the interval . . . or not in the interval.  
 Then for a given level of confidence:

If  $(-3, 2)$ , then "...the difference in population means is between -3 and 2..."

If  $(-7, -4)$ , then "...population mean 1 is 4 to 7 units less than population mean 2..."

If the interval is entirely negative, then we can conclude that  $\mu_1$  is less than  $\mu_2$  by the amount given by the interval with whatever level of confidence was used. Notice that the negative signs are not needed in the written interpretation because they have been interpreted by the word "less" in the comparison.

## Interpreting CI's for a Difference: $\mu_1 - \mu_2 = ?$

The importance of 0 in the interval . . . or not in the interval.

Then for a given level of confidence:

If  $(-3, 2)$ , then "... the difference in population means is between -3 and 2..."

If  $(-7, -4)$ , then "... population mean 1 is 4 to 7 units less than population mean 2..."

If  $(5, 9)$ , then "... population mean 1 is 5 to 9 units more than population mean 2..."

## Two-Sample t-Procedures for Means

## └ Interpreting CI's for a Difference:

 $\mu_1 - \mu_2 = ?$ Interpreting CI's for a Difference:  $\mu_1 - \mu_2 = ?$ 

The importance of 0 in the interval . . . or not in the interval.

Then for a given level of confidence:

If  $(-3, 2)$ , then "...the difference in population means is between -3 and 2..."If  $(-7, -4)$ , then "...population mean 1 is 4 to 7 units less than population mean 2..."If  $(5, 9)$ , then "...population mean 1 is 5 to 9 units more than population mean 2..."

If the interval falls entirely in the positive part of the number line, then it can be concluded that  $\mu_1$  is greater than  $\mu_2$  by the amount given by the interval with whatever level of confidence was used. It makes sense in the order of subtraction that when the answer is positive the first number is bigger than the second number.

# Assessing Conditions with Other Hypothesis Tests

- ▶ Shapiro-Wilk test for normality
- ▶ Levene test for equal variance

## Two-Sample t-Procedures for Means

### └ Assessing Conditions with Other Hypothesis Tests

- Shapiro-Wilk test for normality
- Levene test for equal variance

We'll be introducing two additional hypothesis tests this week: the Shapiro-Wilk test for normality and the Levene test for equal variance.

Both the pooled and separate-variance t-procedures require the samples to be independent and randomly selected as well as normally distributed.

The Shapiro-Wilk test will be used to assess the normality of the samples in addition to looking at boxplots, histograms, and normal probability plots.

Levene's test will be used to determine if the condition of equal population variances is met for the pooled t-interval and test.

# Shapiro-Wilk Test

$H_0$ : The sample was drawn from a normally distributed population.

$H_a$ : The sample was not drawn from a normally distributed population.

## Two-Sample t-Procedures for Means

### └ Shapiro-Wilk Test

#### Shapiro-Wilk Test

$H_0$ : The sample was drawn from a normally distributed population.

$H_a$ : The sample was not drawn from a normally distributed population.

While we won't formally be teaching the mathematical details of the Shapiro-Wilk test and it isn't presented in the textbooks we're using, it is easily performed in R and it is important to understand what is being tested.

The null hypothesis is simply that a particular random sample was drawn from a normally distributed population, while the alternative hypothesis is that it was not.

# Levene Test

$H_0$ : Population variances are equal.

$H_a$ : Population variances are not equal.



## Two-Sample t-Procedures for Means

### └ Levene Test

#### Levene Test

$H_0$ : Population variances are equal.

$H_a$ : Population variances are not equal.

Similarly, Levene's is easily performed in R and it is important to understand what is being tested here as well.

The null hypothesis is simply that the population variances are equal, while the alternative hypothesis is that they are not.

To introduce some terminology here, it is not uncommon to refer to equal population variance as

# Levene Test

$H_0$ : Homogeneity of variance

$H_a$ : Population variances are not equal.

# Two-Sample t-Procedures for Means

## └ Levene Test

### Levene Test

$H_0$ : Homogeneity of variance

$H_a$ : Population variances are not equal.

“homogeneity of variance” and unequal population variances as

# Levene Test

$H_0$ : Homogeneity of variance

$H_a$ : Heterogeneity of variance

# Two-Sample t-Procedures for Means

## └ Levene Test

### Levene Test

$H_0$ : Homogeneity of variance

$H_a$ : Heterogeneity of variance

“heterogeneity of variance” . . . Homogeneity of variance is also called

# Levene Test

$H_0$ : Homoscedasticity

$H_a$ : Heterogeneity of variance

# Two-Sample t-Procedures for Means

## └ Levene Test

### Levene Test

$H_0$ : Homoscedasticity

$H_a$ : Heterogeneity of variance

homoscedasticity and heterogeneity of variance is known as

# Levene Test

$H_0$ : Homoscedasticity

$H_a$ : Heretoscedasticity



## Two-Sample t-Procedures for Means

## └ Levene Test

## Levene Test

 $H_0$ : Homoscedasticity $H_a$ : Heteroscedasticity

heteroscedasticity . . . the hypotheses are probably most commonly written using the

# Levene Test

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

## Two-Sample t-Procedures for Means

## └ Levene Test

Levene Test

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

mathematical symbols that represent the population variances.

## Example

Suppose a national car rental agency wants to compare the average number of miles their customers put on luxury car rentals versus sports car rentals and estimate their difference. The company selects independent, random samples of 15 luxury car rentals and 15 sports car rentals and record the daily mileage for each.

## Two-Sample t-Procedures for Means

### └ Example

#### Example

Suppose a national car rental agency wants to compare the average number of miles their customers put on luxury car rentals versus sports car rentals and estimate their difference. The company selects independent, random samples of 15 luxury car rentals and 15 sports car rentals and record the daily mileage for each.

When reading a problem scenario like this, it is very important for us to be able to pick out the key pieces of information. For example, we should be able to identify the two populations of interest.

Here, we see that there is some national car rental agency that has both luxury car rentals and sports car rentals. And we read that they want to compare the average number of miles for each – this is the response variable.

It also says that the company has selected independent random samples. This is important because these are conditions for both the separate-variance t-procedures and the pooled t-procedures. If these conditions were not clearly stated, we would be in the position of having to assume them in order to proceed.

## Example

Conduct a hypothesis test at a 10% level of significance to compare the population means miles per day for luxury car and sports car rentals. Also construct a 90% confidence interval to estimate the difference in population means.

# Two-Sample t-Procedures for Means

## └ Example

### Example

Conduct a hypothesis test at a 10% level of significance to compare the population means miles per day for luxury car and sports car rentals. Also construct a 90% confidence interval to estimate the difference in population means.

There is no audio for this slide.

# Sample Statistics

```
mean(luxury_cars); sd(luxury_cars)
```

```
## [1] 36.4
```

```
## [1] 10.95967
```

```
mean(sports_cars); sd(sports_cars)
```

```
## [1] 54.2
```

```
## [1] 12.65024
```



## Two-Sample t-Procedures for Means

### └ Sample Statistics

#### Sample Statistics

```
mean(luxury_cars); sd(luxury_cars)

## [1] 36.4
## [1] 10.95967

mean(sports_cars); sd(sports_cars)

## [1] 54.2
## [1] 12.65024
```

Before doing any statistical inference, it is always a good idea to examine the sample statistics. The R code for generating the sample means and standard deviations is shown as well as the output that is generated by each.

We'll assume that we have already loaded the data into R and that the names `luxury cars` and `sports cars` identify vectors that contain each sample.

Clearly the sample mean for the sports cars is higher, but that doesn't necessarily mean that the population means are different. However, with this in mind, if we do find a significant difference in population means, it would stand to reason that the sample with the higher mean would correspond to population with the higher mean.

## Assessing Conditions: Equal Variance (numerically)

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.1.3
```

```
leveneTest(mileage~cartype)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group contains  
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
```

```
##           Df F value Pr(>F)
```

```
## group    1  0.2359 0.6309
```

```
##           28
```

## Two-Sample t-Procedures for Means

### └ Assessing Conditions: Equal Variance (numerically)

#### Assessing Conditions: Equal Variance (numerically)

```
library(car)

## Warning: package 'car' was built under R version 3.1.3

leveneTest(mileage~carmodel)

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group  1  0.2359 0.6309
##      28
```

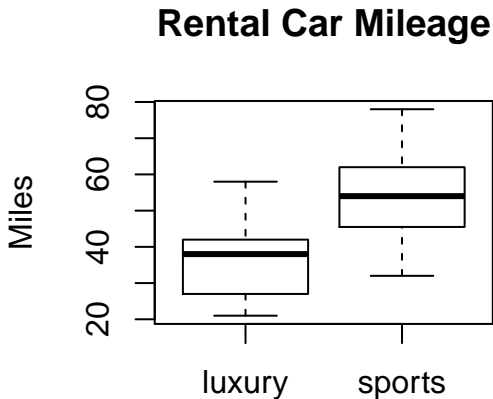
To conduct Levene's test for equal variance in R, the package “car” must be installed and called from the library. (by the way it is purely a coincidence that this example is about cars and this package is called car)

A warning message is shown alerting us that the car package was built under a different version of R than the one we are currently running. Many times this will not affect the output – such as in this case.

Another warning is given when the Levene test is executed, but again, it is more informational rather than something that will affect our output.

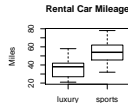
What we are most interested in with this output is the P-value

## Assessing Conditions: Normality (visually)



## Two-Sample t-Procedures for Means

### Assessing Conditions: Normality (visually)

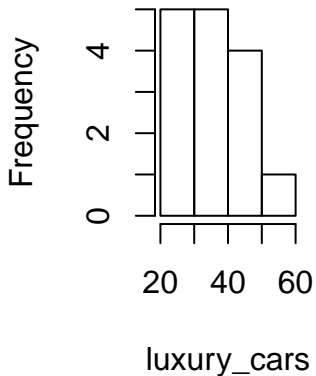


To assess the normality of the samples it is wise to make a visual assessment of boxplots and histograms in addition to conducting a goodness-of-fit test such as the Shapiro-Wilk test for normality.

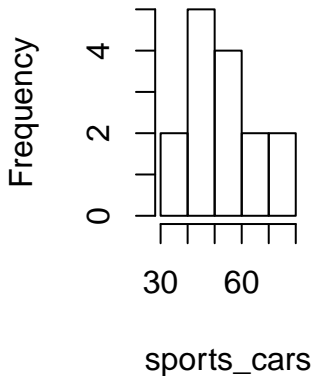
No outliers are seen in these boxplots and both boxplots are relatively symmetric. Once again, however, we can see that the mileage for sports car rentals tends to be higher in the sample, though we have yet to compare population means formally.

# Assessing Conditions: Normality (visually)

**Luxury Cars**



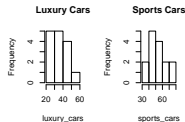
**Sports Cars**



## Two-Sample t-Procedures for Means

### Assessing Conditions: Normality (visually)

Assessing Conditions: Normality (visually)



The histograms are a little blocky and may have a very slight skewness to the right. However, we should keep in mind that the samples contain random variation simply from the process of random sampling and that they are also relatively small samples.

We should not be too hasty to declare them to be non-normal. We would only be skeptical normality if we see heavy skewness, extreme outliers, or a large number of outliers.

## Assessing Conditions: Normality (numerically)

```
shapiro.test(luxury_cars)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  luxury_cars  
## W = 0.9569, p-value = 0.6395
```

```
shapiro.test(sports_cars)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  sports_cars  
## W = 0.9795, p-value = 0.966
```



## Two-Sample t-Procedures for Means

### └ Assessing Conditions: Normality (numerically)

#### Assessing Conditions: Normality (numerically)

```
shapiro.test(luxury_cars)

##
## Shapiro-Wilk normality test
##
## data:  luxury_cars
## W = 0.9569, p-value = 0.6395

shapiro.test(sports_cars)

##
## Shapiro-Wilk normality test
##
## data:  sports_cars
## W = 0.9795, p-value = 0.966
```

When assessing the condition numerically that both samples come from normally distributed populations, the Shapiro-Wilk test is applied to each sample individually.

A p-value of 0.6395 for the luxury cars is large enough that normality for the daily mileage for the population of all luxury cars at this company would not be rejected at a 5% level of significance, and similarly the large p-value of .966 for the sports cars indicates that it would not be unreasonable at all to consider the population of mileage for all sports cars to be normally distributed as well.

## Define the Parameters

Let  $\mu_1$  represent the population mean miles per day that customers put on all luxury car rentals.

Let  $\mu_2$  represent the population mean miles per day that customers put on all sports car rentals.

## Two-Sample t-Procedures for Means

### └ Define the Parameters

#### Define the Parameters

Let  $\mu_1$  represent the population mean miles per day that customers put on all luxury car rentals.

Let  $\mu_2$  represent the population mean miles per day that customers put on all sports car rentals.

Since we were given in the problem statement that the samples were independent and randomly selected and we have determined that the underlying populations are normally distributed with equal variance, we can proceed with either the separate variance t-test or the pooled t-test.

In either case, we should begin by being clear about defining the mathematical symbols we use and identifying the populations for the parameters we are comparing.

# State the Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

## Two-Sample t-Procedures for Means

## └ State the Hypotheses

State the Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

Every hypothesis test must begin with a null and alternative hypothesis. Since the problem statement did not specify a particular direction to test, that is, it did not say “test that the mean for sports cars is higher than for luxury cars – it simply said to compare the population means, so we choose the two-tailed alternative, which states that the population means are not equal to each other.

## R Code and Output

```
t.test(mileage~cartype,data=cardata,var.equal=TRUE,conf.level=0.90)
```

```
##  
## Two Sample t-test  
##  
## data: mileage by cartype  
## t = -4.1188, df = 28, p-value = 0.0003053  
## alternative hypothesis: true difference in means is not equal to 0  
## 90 percent confidence interval:  
## -25.1516 -10.4484  
## sample estimates:  
## mean in group luxury mean in group sports  
## 36.4 54.2
```

## Two-Sample t-Procedures for Means

### └ R Code and Output

#### R Code and Output

```
t.test(mileage~cartype,data=cardata,var.equal=TRUE,conf.level=0.90)

##
## Two Sample t-test
##
## data: mileage by cartype
## t = -4.1188, df = 28, p-value = 0.0003053
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -25.1516 -10.4484
## sample estimates:
## mean in group luxury mean in group sports
##                36.4                54.2
```

In order to get the t-test and t-interval output using R the `t.test` function is used. I'll show you the structure of the data and the variable names in just a few more slides.

Since the pooled t-test conditions are met and it is an exact procedure, I have chosen it to compare the population means, as indicated by the option `var.equal=TRUE` within the `t.test` function.

Since a 90% confidence interval was requested, I also had to specify the option `conf.level=0.90`. If this was not specified, the function uses `.95` as the default level of confidence.

The test statistic, degrees of freedom, and p-value for the two-sample t-test are given as well as a note indicating the default option of the two-tailed alternative that the difference in population

# Test Conclusions

Statistical conclusion: Reject  $H_0$  at  $\alpha = 0.10$ .

Interpretation: There is sufficient evidence to claim that the population mean miles per day customers put on all of this company's luxury cars is different from the population mean miles per day customers put on all of their sports cars ( $P = 0.0003$ ).



## Two-Sample t-Procedures for Means

### └ Test Conclusions

#### Test Conclusions

Statistical conclusion: Reject  $H_0$  at  $\alpha = 0.10$ .

Interpretation: There is sufficient evidence to claim that the population mean miles per day customers put on all of this company's luxury cars is different from the population mean miles per day customers put on all of their sports cars ( $P = 0.0003$ ).

Hypothesis test conclusions can be written as a statistical conclusion, simply stating whether or not the null hypothesis was rejected and reporting the level of significance. Here  $H_0$  is rejected at a 10% level of significance since the p-value, 0.0003, is less than 0.10.

In order to clearly communicate the result of the hypothesis, a practical conclusion, or interpretation, is also given in the context of the problem. Please take a moment to read the interpretation and go back and compare it to the wording in the original problem statement. Stating the p-value in parenthesis is commonly done in research publications.

# Interpreting the Confidence Interval

With 90% confidence, the population mean mileage per day on all luxury cars is 10.45 to 25.15 miles less than for all sports cars for this rental company.

## Two-Sample t-Procedures for Means

### └ Interpreting the Confidence Interval

#### Interpreting the Confidence Interval

With 90% confidence, the population mean mileage per day on all luxury cars is 10.45 to 25.15 miles less than for all sports cars for this rental company.

A typical confidence interval interpretation is shown here. Note that the inference is in reference to population means, not sample means. Also note that negative signs are not brought into the interpretation with the numerical values of the lower and upper bounds of the confidence interval, but rather have been interpreted in that we are claiming that the population mean mileage per day for luxury cars is less than for sports car rentals for this company.

Since we know the order of subtraction was luxury car minus sports car, and we see the interval in the output from R that is entirely negative, we can conclude that the population mean for luxury cars must be less than the mean for sports cars by somewhere between 10.45 and 25.15 miles, that is, with 90% confidence.

## Structuring the Data in R

```
mileage <- c(50,58,43,48,50,66,32,73,55,78,43,66,57,40,54,  
            26,33,21,37,39,23,48,58,28,39,22,38,41,50,43);  
cartype <- c(rep("sports",15),rep("luxury",15));  
cardata <- data.frame(mileage,cartype)  
sports_cars <- mileage[cartype=="sports"]  
luxury_cars <- mileage[cartype=="luxury"]
```

## Two-Sample t-Procedures for Means

### └ Structuring the Data in R

#### Structuring the Data in R

```
mileage <- c(50,58,43,48,50,66,32,73,55,78,43,66,57,40,54,  
            26,33,21,37,39,23,42,58,28,39,22,38,41,50,43);  
cartype <- c(rep("sports",15),rep("luxury",15));  
cardata <- data.frame(mileage,cartype)  
sports_cars <- mileage[cartype=="sports"]  
luxury_cars <- mileage[cartype=="luxury"]
```

Though it wasn't shown in the problem statement, the data for this example would have been provided in some way. If it was simply printed in a book and it was up to me to get it into R, I would have to create a data frame as shown here.

In the data frame, which I called cardata, the data are entered and stored in two vectors, one called mileage and one called cartype. Mileage stores the 30 observed miles per day of all the cars in the sample with the 15 sports car mileage values entered first followed by the 15 luxury car values. Think of a matrix with 2 columns and 30 rows.

Cartype stores a categorical indicator of which type of rental car, thus identifying which population each sample value was selected