R Part IV - Statistics *t*-tests

Dr. Irene Vrbik

University of British Columbia Okanagan

irene.vrbik@ubc.ca

Term 2, 2018W

Introduction

- Today we will dive a little deeper into some of the statistical tools you may find yourself using in R.

- While these tests are not specific to R, the specific syntax used throughout this lecture will be.

- As always, you can find the code used throughout this lecture on Canvas.

- More specifically we will be looking at the `t.test()` function in R for the purpose of hypothesis testing.

- We will focus on
  1. comparing one-sample mean to a to a stipulated value,
  2. comparing the means of two independent groups:
  3. comparing the means of paired samples:

Hypothesis Testing

Hypothesis testing is an essential procedure in statistics.

Hypothesis tests are used in virtually every field of study, here are some applications to name just a couple:

1. in determining whether a value is below/above a certain threshold, eg. test if the average cadmium level in edible mushrooms is below the tolerable value of 3 mg/kg
2. in determining an effect in controlled experiments, eg. to compare a new medical treatment to a placebo.
3. in determining effectiveness of marketing, eg. is this years sales better than the previous year?

Background

Hypothesis testing is form of inferential statistics.

- **Statistical inference** involves forming judgements on a population of interest based on a random sample drawn from that population.
- In contrast to descriptive statitistcs—which does not allow us to make conclusions beyond the data we have observed—inferential statitics aims at using information provided by the sample to infer and make predictions about a population from which it came.

Background

- The general family of t-tests refer to statistical hypothesis tests which rely on the bell-shaped t-distribution.

- While the complete collection of t-tests along with their mathematical justification are beyond the score of this course, we highlight a commonly used t-test for performing inference on population *means*.

One-sample t-test

Assumptions

There are assumptions that need to be met before performing t-test.

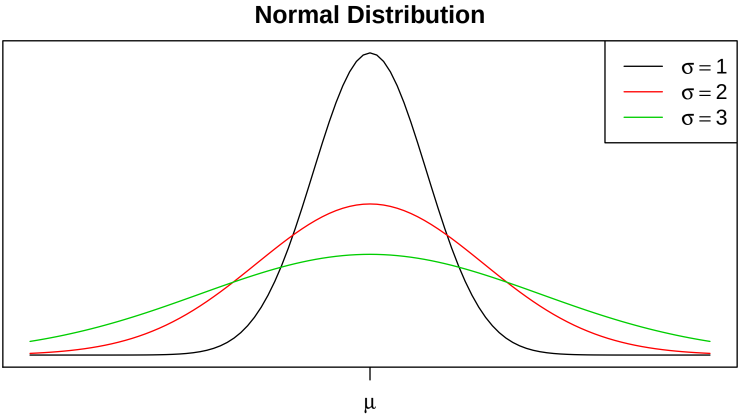
1. Population of interest is normally distributed.
2. Independent random samples are taken.

While there are statistical methods for testing assumptions (eg. Shapiro-Wilk test for normality), for breverity, we assume that they have been met.

Normal distribution

- The normal distribuiton is the most widely used perhaps the most well recognizable distribution by it's "bell-shape".
- The normal distribution is characterized by two parameters: its mean (μ) and standard deviation (σ).
  - the **mean** provides the location of the bell's center
  - and the **standard deviation** describes how spread out, or 'fat', that bell shape is.

The Normal Distribution



The tails of this distribution actually run from -∞ to ∞.

Null Hypothesis

- For the one sample problem, we may wish to test the hypothesis that the population is centered at some supposed value μ_0.

- In more statistical terms, we may wish to test the **null hypothesis** of $H_0 : \mu = \mu_0$.

- In this course the null hypothesis, H_0, always contains a statement of no change ($=$). *This is not universal across textbooks.

Alternative Hypothesis

- H_0 is tested against a competing statement called the **alternative hypothesis** H_1 (sometimes written H_A).
- We can either make this alternative one or two-sided:

  $H_1 : \mu \neq \mu_0$ (two-sided)

  $H_1 : \mu < \mu_0$ (one-sided, lower-tailed)

  $H_1 : \mu > \mu_0$ (one-sided, upper-tailed)
- H_0 and H_1 should always be written in terms of the *population* parameters (in this case μ).

Alternative Hypothesis

- The direction of our alternative hypothesis will depend on the situation at hand.
- A two-sided alternative may be preferred if a deviation in either direction is just as grave.
  - eg. to compare a new medical treatment to a placebo (we care both if the treatment is effective and if it is harmful).

Alternative Hypothesis

- One-sided alternatives may be preferred if results of your test are only relavent in one direction.
  - eg. is this years sales better than the previous year? (if yes, employees get a bonus, if not, nothing happens)
  - eg. do less than half the adults in a certain area favour the construction of an outdoor rink?

One Sample Test Example

- This test can be used to determine if the sample mean $\overline{x}$ is *significantly* different then some hypothesized value μ_0.
- Note that with any sample, it would be rare that a sample mean be *exactly* equal to the hypothesized value.
    - eg. even if the average height of men on campus is 69.3 in, a sample of 50 students may yeild a sample mean of 68.9 in.
- The question therefore becomes: how far does my sample mean need to stray from the hypothesized value before I begin to question it's validity?

One Sample Test: Calculate Test Statistic

To answer this question we use a **test statistic** which follows some known distribution.

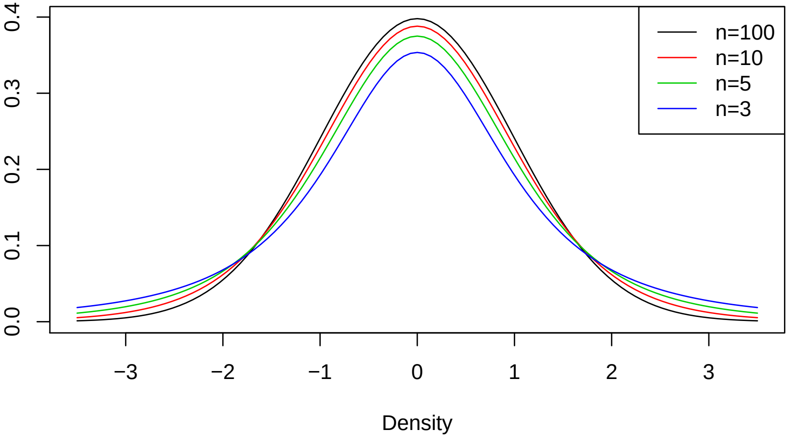
For the one sample test the *t*-test statistic is calculated as:

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{1}$$

where $\overline{x}$ is the sample mean, s is the sample standard deviation, n is the sample size, and μ_0 is the hypothesized mean value.

As alluded to earlier, this statistic follows a *t*-distribution.

The distribution of (1) looks like this:



It's shape is determined by the degrees of freedom parameter $\nu = n - 1$ (the smaller n is, the fatter/more squished it appears)

One Sample Test Example

- Consider the `car_data.csv` file uploaded to Canvas.

```r
car_data <- read.csv("data/car_data.csv")
```

- This data contains information on 8 variables from 30 cars.
- We would like to test if the average mileage (`km.L`) differs from 10km/L.

Step 1: Hypotheses Statements

- The first step is always to write the null and alternative hypothesis:

  $$H_0 : \mu = 10 \quad vs \qquad\qquad H_1 : \mu \neq 10$$

  (i.e. our hypothesized value is $\mu_0 = 10$).
- Analogous to a courtroom, we believe H_0 is true until evidence (in the form of data) suggests otherwise.

Step 2: calculate test statitic

- Step 2 is to calculate our test statistic:

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

N.B. once we plug in values to our test statistic, we call it the *observed* test statistic t_{obs}

$$t_{obs} = \frac{10.3553333 - 10}{\frac{1.210354}{\sqrt{30}}} = 1.6079931$$

Step 2: calculate test statitic

```
mu0 = 10 # hypothesize value
n = nrow(car_data)
tobs = (mean(car_data$km.L)- mu0)/(sd(car_data$km.L)/sqrt(n))
tobs

## [1] 1.607993
```

Step 3: convert to a *p*-value

- Step 3 is to convert this test statistic to a *p*-value.
- A *p*-value is calculated as:

$$2 * P(t >\mid t_{obs} \mid) \text{ if } H_1 : \mu \neq \mu_0 \text{ (two-sided)}$$

$$P(t < t_{obs}) \text{ if } H_1 : \mu < \mu_0 \text{ (one-sided, lower-tailed)}$$

$$P(t > t_{obs}) \text{ if } H_1 : \mu > \mu_0 \text{ (one-sided, upper-tailed)}$$

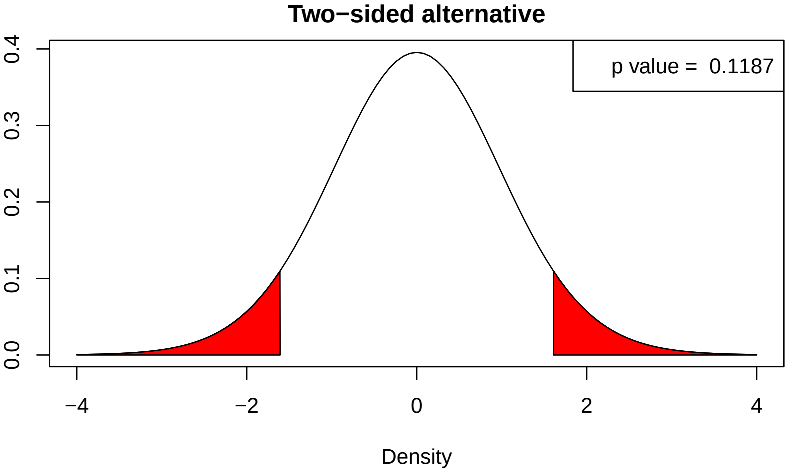
Where

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

is a *t*-distributed random variable with $n - 1$ degrees of freedom.

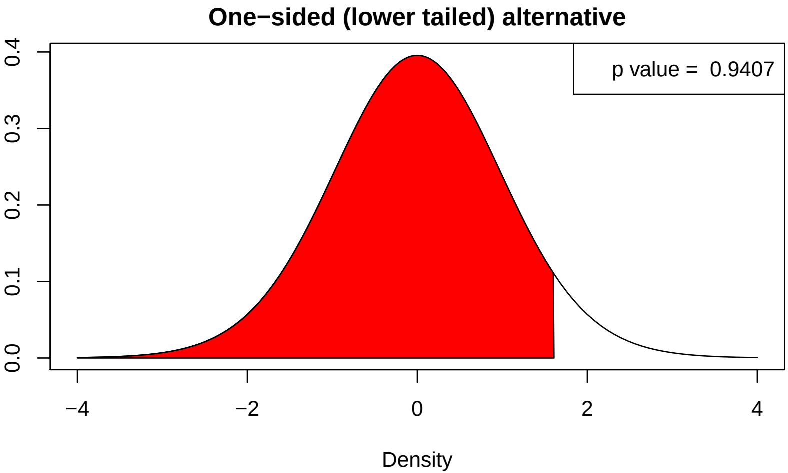
Step 3: convert to a *p*-value

We can visualize this probability as the following area underneath the curve:



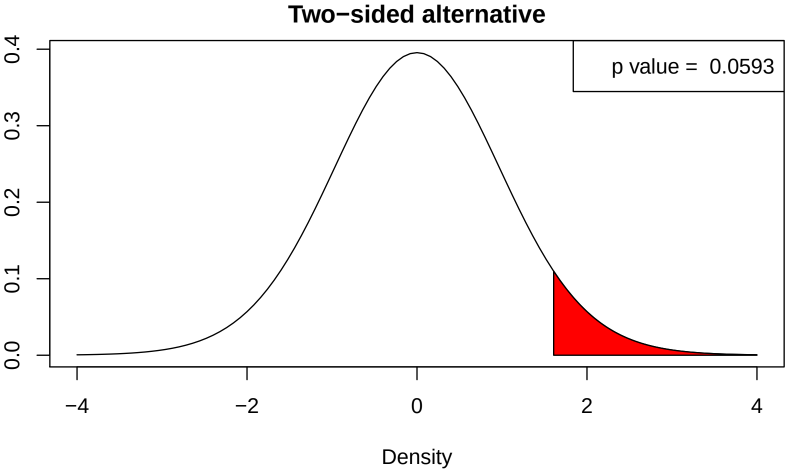
Step 3: convert to a *p*-value

Had we stated the one-sided lower tailed alternative, our p-value would have been



Step 3: convert to a *p*-value

Had we stated the one-sided upper tailed alternative, our p-value would have been



Step 4: State Conclusion

- Based on the evidence (ie data presented by the sample) we must decide whether or not we reject the hyptohesis or not.

- This decision depends on our **confidence level**, often expressed as 1-α, with α begin the **significance level**.

- The standard confidence level is 0.95 (ie. $\alpha = 0.05$)[1]

$$\text{Conclusion} = \begin{cases} \text{reject } H_0 & \text{if } p\text{-value } \leq \alpha \\ \text{fail to reject } H_0 & \text{if } p\text{-value } > \alpha \end{cases}$$

[1] unless stated otherwise, you may always assume $\alpha = 0.05$

A note on stating conclusion

- Unless we have a complete sensus of the entire population we can never say that a given null hypothesis is true.
- Our conclusion for our car example may be written:

  Since the p-value (=0.1187) is greater than the significance level ($\alpha = 0.05$), we fail to reject H_0. Therefore, there is insufficient evidence to suggest that the average car mileage differs from 10.
- Notice how we add context to this decision by referencing the alternative hypothesis in the context of the original problem.

Decision and Conclusion (using *P*-value)

More generically,

If *p*-value $> \alpha$, our decision/conclusion would be:

1. Fail to reject the null hypothesis.
2. There is insufficient evidence to suggest that the mean value is less than/greater than/different than the test value (will depend on alternative hypothesis)

Decision and Conclusion (using P-value)

If $p - \text{value} \leq \alpha$, our decision/conclusion would be:

1. Reject the null hypothesis.
2. There is statistically significant evidence to suggest that the mean value is less than that/greater than/ different than the test value (will depend on alternative hypothesis).

A note on the confidence levels

- Loosely speaking we can view α as our adopted risk when performing this test.

- That is to say, if the null hypothesis is true, 5% of the test we conduct under these conditions will result in the conclusion to reject the null hypothesis.

- We refer to this type of mistakes as Type 1 errors.

Hypothesis testing in R

- Now that we have a basic understanding of what these tests are doing, lets code them up in R. The general syntax:

  `t.test(x=mydata,alternative="two.sided",mu =` μ_0
  `,conf.level=0.95,...)`

- The default `alterative` is `"two.sided"` with the options `"less"`, or `"greater"`.
- The default `conf.level` = 0.95 (usually what we want)
- To see the help file on this function, we type `?t.test`.

One Sample Test: Cars Example

The following code will perform a one-sample t-test which tests:

$$H_0 : \mu = 10 \qquad vs \qquad H_1 : \mu \neq 10$$

```r
car_data <- read.csv("data/car_data.csv")# read in the data
mu0 = 10 # hypothesize value
t.test(x=car_data$km.L,             # sample mileage
       alternative="two.sided",  # two-side H_A
       mu=mu0) # set the hypothesized value mu0
# since a two-sided test is the default, we could have typed:
t.test(x=car_data$km.L, mu=mu0)
```

Compare with our t_{obs} and p-value calulated on slide 19 and 22.

```
##
##  One Sample t-test
##
## data:  car_data$km.L
## t = 1.608, df = 29, p-value = 0.1187
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.90338 10.80729
## sample estimates:
## mean of x
##  10.35533
## [1] 8.807341
```

Exercise

1. Redo this test using the lower-tailed alternative hypothesis.

   Verify that you get the same *p*-value as calculated on page 23

2. Redo this test using the upper-tailed alternative hypothesis.

   Verify that you get the same *p*-value as calculated on page 24

Two-sample t-tests

- The one-sample t-test deals with making inference on a single sample for a single population of interest.
- Often we are concerned with *two* samples from potentially different population distributions.
- The goal of a two-sampled t-test is usually to test the hypothesis that two samples have the same mean.

- In more statistical words, if we use μ_A and μ_B to denote the true population mean of group A and group B, respectively, we may want to test whether:

$$H_0 : \mu_A = \mu_B \quad vs \qquad\qquad H_1 : \mu_A \neq \mu_B$$

- An equivalent way of saying this is:

$$H_0 : \mu_A - \mu_B = 0 \quad vs \qquad\qquad H_1 : \mu_A - \mu_B \neq 0$$

- More generically, we could test:

$$H_0 : \mu_d = d_0 \quad vs \qquad\qquad H_1 : \mu_d \neq d_0$$

  where d_0 is our hypothesize value for the μ_d mean *difference* between group A and B.

Two-sample t-tests

- Two-sampled t-test can actually be broken into two categories:
  1. paired
  2. unpaired
- **paired** data occur when we are obtaining two measurements on the *same* individual.
  - Eg. midterm 1 mark and midterm 2 mark for each student in Data 301.
- **unpaired** data occur when we have two *independent* samples.
  - Eg. comparing the heights of men and women

Two-sample *t*-tests

- In either case, we are still interested in comparing the group means.
- The only difference in terms of R-code is that we will need to specify `paired = TRUE` when the data are paired.
- N.B. for unpaired data we can set `paired = FALSE`, but since this is the default setting in R, this specification may be omitted from our code.

Two Sample Unpaired

An unpaired (independent) two sample test compares two independent samples to determine if there is a difference between the groups.

Example

1. Compare effectiveness of two different drugs tested on two sets of patients.
2. Experiment versus control samples.

Two Sample Unpaired Example

Hypothesis Statement

Using the `beaver2` dataset in R, test the hypothesis that there is no difference between the average temperature of active beavers and non-active beavers.

Letting μ_1 and μ_2 represent the mean temperatures of active and non-active beavers, respectively, our hypotheses may be written:

$$H_0 : \mu_1 - \mu_2 = 0 \rightarrow \mu_d = 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 = 0 \rightarrow \mu_d \neq 0$$

N.B. we can change the hypothesized value $\mu_0 = 0$ to any value and the choice of an upper/lower-tailed alternative.

Unpaired two-sample *t*-test

The general syntax for performing an unpaired two-sample t.test:

```
t.test(x=A, y=B,alternative="two.sided",mu = d0
        ,conf.level=0.95, data=mydata, ...)
```

where `A` and `B` contain the samples from group A and B, respectively. Alternatively, we could specify a *formula*

```
t.test(formula=C~D,alternative="two.sided",mu = d0
        ,conf.level=0.95, data=mydata, ...)
```

where `C` is the sample of measurements and `D` is a factor indicating the category to which each sample belongs.

Two Sample Unpaired Example

Lets have a look at the beaver2 data:

```
head(beaver2,3)

##   day time  temp activ
## 1 307  930 36.58     0
## 2 307  940 36.73     0
## 3 307  950 36.93     0

tail(beaver2,3)

##     day time  temp activ
## 98  308  140 38.01     1
## 99  308  150 38.04     1
## 100 308  200 38.07     1
```

Two Sample Unpaired Example

- As seen in the help file `?beaver2` the `activ` variable denotes activity level of the beaver (1 = active, 0=non-active)
- We could create a subset of this data and perform a t-test as follows:

```
beaverA <- subset(beaver2, activ==1) #active beavers
beaverB <- subset(beaver2, activ==0) #non-active beavers
t.test(beaverA$temp, beaverB$temp)
```

To see the more verbose specification:

```r
t.test(beaverA$temp, # temperatures of 62 active beavers
       beaverB$temp, # temps of 38 inactive beavers
       alternative = "two.sided", # H_1: mu_A - mu_B != d_0
       mu = 0,                    # default d_0 = 0
       paired = FALSE,            # default (data unpaired)
       conf.level = 0.95          # default (ie. alpha = 5%)
       )
```

R output

```
##
##  Welch Two Sample t-test
##
## data:  beaverA$temp and beaverB$temp
## t = 18.548, df = 80.852, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  0.7197342 0.8927106
## sample estimates:
## mean of x mean of y
##  37.90306  37.09684
```

Two Sample Unpaired Example

Alternatively we could have used the formula option:

```r
t.test(temp~activ, data=beaver2)
```

This tells R that the relevant data appear in the `temp` column of data.frame `beaver2` and that the these samples should be divided according to the factor `active`.

Both options should produce identical *p*-values.[2]

[2] sign of test statistic may change depending on μ_d ($\mu_A - \mu_B$ or $\mu_B - \mu_A$)

Two Sample Unpaired Example

```
(bsum <- t.test(temp~activ, data=beaver2))

##

##  Welch Two Sample t-test

##

## data:  temp by activ

## t = -18.548, df = 80.852, p-value < 2.2e-16

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

## 95 percent confidence interval:

##  -0.8927106 -0.7197342

## sample estimates:

## mean in group 0 mean in group 1

##        37.09684        37.90306
```

Footnote

- As we are only interested in the *p*-value, you shouldn't be bothered that the test statistic changed signs depending on how we coded this up in R.
- But in case you are wondering, the explicit notation is testing:

$$H_0 : \mu_1 - \mu_0 = 0 \quad vs. \quad H_1 : \mu_1 - \mu_0 \neq 0$$

  whereas the formula notation is testing:

$$H_0 : \mu_0 - \mu_1 = 0 \quad vs. \quad H_1 : \mu_0 - \mu_1 \neq 0$$

- For two-sided alternatives this matters not, but for one-sided tests, we need to pay close attention to the direction of $</>$.

To produce output identical to the formula notation use:

```
# places inactive beavers in the first argument:
(beavertest <- t.test(beaverB$temp,beaverA$temp))

##
##  Welch Two Sample t-test
##
## data:  beaverB$temp and beaverA$temp
## t = -18.548, df = 80.852, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -0.8927106 -0.7197342
## sample estimates:
## mean of x mean of y
##  37.09684  37.90306
```

- Since the *p*-value $\ll \alpha$, there is very strong evidence to suggest that the average temperature between active and non-active beavers are different from one another.

- We use the language of “strong" since the *p*-value is so small.

- Rather than reading the *p*-value from the output table, we can also reference the *p*-value from the `htest` (hypthesis test) object:

```
beavertest$p.value # to see the exact p-value

## [1] 7.269112e-31
```

Two Sample Paired Test

A paired (dependent) two sample test compares two dependent samples to see if there is a difference between the groups.

1. This test typically uses multiple measurements on one subject.
2. Also called a "repeated measures" test.

Examples

1. Affect of treatment on a patient (before and after)
2. Do cars get better mileage with different grades of gasoline?

Paired data

- We can visualize it as follows:

Group 1	Group 2	Difference
x_1	y_1	$d_1 = y_1 - x_1$
x_2	y_2	$d_2 = y_2 - x_2$
$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$d_n = y_n - x_n$

- Notice that Group A and B will neccessarily have the same number of observations. (unpaired test will not necessarily have the same number of observations)

Paired data

- If we are interested in testing if the true mean of Group 1 (μ_1) differs from true mean of Group 2 (μ_2), that is equivalent to testing true mean differences μ_d is significantly different from 0.

- This is equivalent to the one-sample t.test that we introduced first!

- Let's look at example.

Two Sample Paired Test Example

Hypothesis Statement

The `ahtlete.csv` dataset contains data on ten athletes and their speeds for the 100m dash before training (Training = 0) and after (Training = 1).

Test the hypothesis that the training has no effect on the times of the athletes., i.e. test if the mean difference is different than 0.

$$H_0 : \mu_d = 0 \quad vs. \quad H_1 : \mu_d \neq 0$$

where $\mu_d = \mu_{\text{Train=0}} - \mu_{\text{Train=1}}$

`Training` = 1 if the athlete has trained, 0 otherwise.

```
athlete = read.csv("data/athlete.csv", header=TRUE)
head(athlete,3)
##   Athlete Time Training
## 1       1 12.9        0
## 2       2 13.5        0
## 3       3 12.8        0
tail(athlete,3)
##    Athlete Time Training
## 18       8 15.9        1
## 19       9 16.0        1
## 20      10 11.1        1
```

Two Sample Paired Test Example

athlete data

Notice how each group contains the same athletes (indexed by number)

```
noTrain <- subset(athlete, Training==0)
Train <- subset(athlete, Training==1)
# these are performed on the same athletes!
noTrain$Athlete

##  [1]  1  2  3  4  5  6  7  8  9 10

Train$Athlete

##  [1]  1  2  3  4  5  6  7  8  9 10
```

This is a neccessity for paired data!

Notice the only change is that we need to specify paired=TRUE

```r
t.test(Time~Training, data=athlete, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  Time by Training
## t = -0.12031, df = 9, p-value = 0.9069
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -0.5544647  0.4984647
## sample estimates:
## mean of the differences
##                  -0.028
```

Alternative way of coding:

```
t.test(noTrain$Time, Train$Time, data=athlete, paired=TRUE)

##
##  Paired t-test
##
## data:  noTrain$Time and Train$Time
## t = -0.12031, df = 9, p-value = 0.9069
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -0.5544647  0.4984647
## sample estimates:
## mean of the differences
##                  -0.028
```

Notice how this is the same as the following one-sample *t*-test:

```
ds <- noTrain$Time - Train$Time
t.test(ds)
```

```
##
##  One Sample t-test
##
## data:  ds
## t = -0.12031, df = 9, p-value = 0.9069
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.5544647  0.4984647
## sample estimates:
## mean of x
##    -0.028
```

Two Sample Paired Test Example

Decision and Conclusion

- As our pvalue = 0.9069 is larger than our significance level of 0.05, we fail to reject the null hypothesis.
- Hence there is insufficient evidence to suggest that there is a difference between pre and post training times.
- Notice that we obtain a different *p*-value when we neglect to specify that the data is paired ...
- While this made no difference on the conclusion for this example, it could make a big difference in other practical applications.

```
##
##  Welch Two Sample t-test
##
## data:  Time by Training
## t = -0.024091, df = 17.726, p-value = 0.981
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -2.472494  2.416494
## sample estimates:
## mean in group 0 mean in group 1
##          14.502          14.530
```

Question

How many of the following are true?

1. There is no difference between paired and unpaired *t*-tests.
2. Unpaired *t*-test must have the same number of observations in each group
3. The direction of the alternative hypothesis (ie. $>$, $<$, $\neq$) will have no affect on the calculated *p*-value
4. If the calculated *p*-value is greater than our significance level α we conclude that H_0 is true.

A) 0 B) 1 C) 2 D) 3 E) 4

Question

How many of the following are true?

1. There is no difference between paired and unpaired t-tests. ✗
2. Unpaired t-test must have the same number of observations in each group
3. The direction of the alternative hypothesis (ie. $>$, $<$, $\neq$) will have no affect on the calculated p-value
4. If the calculated p-value is greater than our significance level α we conclude that H_0 is true.

A) 0 B) 1 C) 2 D) 3 E) 4

Question

How many of the following are true?

1. There is no difference between paired and unpaired *t*-tests. ✗
2. Unpaired *t*-test must have the same number of observations in each group ✗
3. The direction of the alternative hypothesis (ie. $>$, $<$, $\neq$) will have no affect on the calculated *p*-value
4. If the calculated *p*-value is greater than our significance level α we conclude that H_0 is true.

A) 0 B) 1 C) 2 D) 3 E) 4

Question

How many of the following are true?

1. There is no difference between paired and unpaired *t*-tests. ✗
2. Unpaired *t*-test must have the same number of observations in each group ✗
3. The direction of the alternative hypothesis (ie. $>$, $<$, $\neq$) will have no affect on the calculated *p*-value✗
4. If the calculated *p*-value is greater than our significance level α we conclude that H_0 is true.

A) 0 B) 1 C) 2 D) 3 E) 4

Question

How many of the following are true?

1. There is no difference between paired and unpaired t-tests. ✗
2. Unpaired t-test must have the same number of observations in each group ✗
3. The direction of the alternative hypothesis (ie. $>$, $<$, $\neq$) will have no affect on the calculated p-value✗
4. If the calculated p-value is greater than our significance level α we conclude that H_0 is true.✗

A) ***0*** B) 1 C) 2 D) 3 E) 4

Which is the most appropriate test for each of the following?

Question

Is the average final grade for Data 301 greater than 70%?

A) One-sampled *t*-test

B) Two-sample *t*-test (unpaired)

C) Two-sample *t*-test (paired)

D) None of the above

Which is the most appropriate test for each of the following?

Question

Is the average final grade for Data 301 greater than 70%?

A) ***One-sampled t-test***

B) Two-sample t-test (unpaired)

C) Two-sample t-test (paired)

D) None of the above

Which is the most appropriate test for each of the following?

Question

Does a student's mark improve after studying (measurements taken on same student)?

A) One-sampled t-test

B) Two-sample t-test (unpaired)

C) Two-sample t-test (paired)

D) None of the above

Which is the most appropriate test for each of the following?

Question

Does a student's mark improve after studying (measurements taken on same student)?

A) One-sampled t-test

B) Two-sample t-test (unpaired)

C) ***Two-sample t-test (paired)***

D) None of the above

Which is the most appropriate test for each of the following?

Question

Has the average student height at UBCO increased since 1990?

A) One-sampled t-test

B) Two-sample t-test (unpaired)

C) Two-sample t-test (paired)

D) None of the above

Which is the most appropriate test for each of the following?

Question

Has the average student height at UBCO increased since 1990?

A) One-sampled t-test

B) ***Two-sample t-test (unpaired)***

C) Two-sample t-test (paired)

D) None of the above

Which is the most appropriate test for each of the following?

Question

Does radiation reduce the size of tumours when used to treat patients? Measurements were taken before and after treatment.

A) One-sampled t-test

B) Two-sample t-test (unpaired)

C) Two-sample t-test (paired)

D) None of the above

Which is the most appropriate test for each of the following?

Question

Does radiation reduce the size of tumours when used to treat patients? Measurements were taken before and after treatment.

A) One-sampled *t*-test

B) Two-sample *t*-test (unpaired)

C) ***Two-sample t-test (paired)***

D) None of the above

Which is the most appropriate test for each of the following?

Question

Are college graduates better than high school graduates at standardized tests?

A) One-sampled t-test

B) Two-sample t-test (unpaired)

C) Two-sample t-test (paired)

D) None of the above

Which is the most appropriate test for each of the following?

Question

Are college graduates better than high school graduates at standardized tests?

A) One-sampled t-test

B) ***Two-sample t-test (unpaired)***

C) Two-sample t-test (paired)

D) None of the above

Question

Using the car data, test the hypothesis that the mean distance traveled (`Distance`) at each fill up is less than 450 miles.

Question

Use the car data to see if the mean distance for 2015 fill ups is different than the mean distance for 2016 fill ups.

- Another important statistical method is least squares regression.
- We have already seen how to fit a linear regression model to data in Excel and Python.
- The remainder of this lecture demonstrates how to perform basic regression analyses in R.

Linear models in R

Recall that a linear model is an equation that relates a response variable (y) to some explanatory variables (x's). The general form of the model is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \tag{2}$$

Linear models in R

Even if a linear relationship exists between the x's and y, due to the natural variability that we experience in the real world, we might expect observations to fall randomly around some close proximity of (2).

We model this using:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdot + b_n x_{bi} + \varepsilon_i$$

where ε_i denotes the error term associated with observation i.

Linear models in R

Assumptions

We are assuming

1. Residuals are independent.
2. Residuals are normally distributed.
3. Residuals have a mean of 0 for all X.
4. Residuals have constant variance.

Fitting Regression Models

- In R, the general syntax for fitting a regression model:

```
lm(formula, data, subset,...)
```

- `formula:` a symbolic description of the model to be fitted
- `subset:` an optional vector specifying a subset of observations to be used in the fitting process.

Fitting Regression Models

- Using the `car` data, lets fit a linear model which attempts to explain the response variable `km.L` (ie mileage) by the the size of the tank (`Litres`) and the distance traveled at each fill-up (`Distance`).

```
model <- lm(km.L~Litres+Distance, data=car_data)
model
```

```
##
## Call:
## lm(formula = km.L ~ Litres + Distance, data = car_data)
##
## Coefficients:
## (Intercept)       Litres     Distance
##    10.35447     -0.33295      0.03251
```

```
model$coefficients

## (Intercept)      Litres    Distance
## 10.35447098 -0.33294847  0.03250697
```

The formula can be then be created using the values stored in `model$coefficients`

`Km.L = 10.35447 -0.33295*Litres + 0.03251*Distance`

We can see a more verbose output by calling the `summary()` function on the output of the `lm` command:

```r
summary(model)
```

```
##
## Call:
## lm(formula = km.L ~ Litres + Distance, data = car_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80592 -0.14334 -0.03808  0.14876  0.59345
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.354471   0.247562   41.83   <2e-16 ***
## Litres      -0.332948   0.016514  -20.16   <2e-16 ***
## Distance     0.032507   0.001743   18.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3129 on 27 degrees of freedom
## Multiple R-squared:  0.9378,Adjusted R-squared:  0.9332
## F-statistic: 203.5 on 2 and 27 DF,  p-value: < 2.2e-16
```

We can obtain things like residuals, R-squared values very easily:

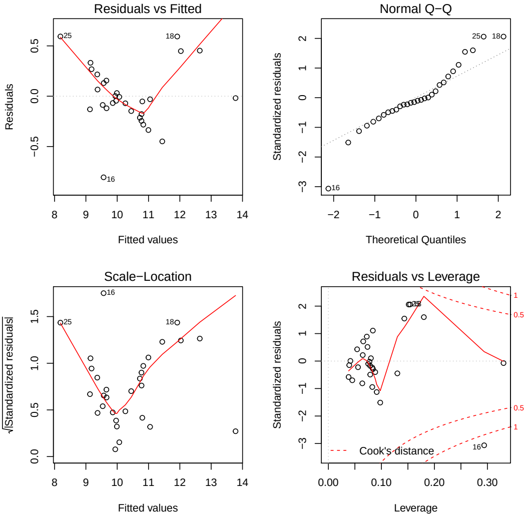
```
e <- residuals(model)
head(e)

##            1            2            3            4            5
## -0.449076892 -0.030541111 -0.051871604  0.332066493  0.001860779
##            6
## -0.245187648

smod <- summary(model)
smod$r.squared

## [1] 0.9377844
```

```
par(mfrow=c(2,2)); # divides plot window into a 2 x 2 matrix
plot(model) # some useful diagnostic plots:
```



Conclusion

- Like much of what we cover in this course, the materials reviewed here are to expose you the available tools for data analysis rather than provide all of the “under the hood" details.
- If you would like to learn more about the justifications behind using these statistical methods, I would suggest DATA 101 (more time and discussion on R and linear regression) STAT 230 (more discussions on statistical inference and hypothesis testing) and DATA 311 (machine learning).