R Part III

Dr. Irene Vrbik

University of British Columbia Okanagan

`irene.vrbik@ubc.ca`

Term 2, 2018W

Introduction

- Now that we covered the fundamental concepts of programing in R, we can discuss some the important statistical analyses that are commonly performed with this software (and others).

- Some of these methods we would have seen already in our Excel and Python unit, others will be seeing for the first time.

- 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.

- We begin with a basic statistical test for comparing continuous data.

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

- We will focus on the one-sample problem which compares a sample to a stipulated value, and two-sample problem for comparing two groups.

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 few:

1. in determining an effect in controlled experiments, eg. to compare a new medical treatment to a placebo.

2. in determining effectiveness of marketing, eg. is this years sales better than the previous year?

3. eg. do less than half the adults in a certain area favour the construction of an outdoor rink?

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, 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

Normal Distribution
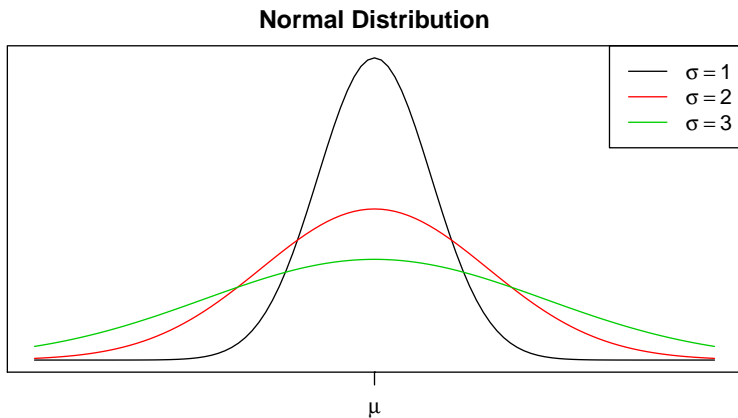
The tails of this distribution actually run from $-\infty$ to ∞.

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

- A one sample test can be used to compare a sample to a model or known population/estimate.

- To put another way, this test is used to determine if the sample mean $\overline{x}$ is significantly different then some hypothesized value μ_0.

- As an example, using the car data test if the average mileage is different than 10km/L.

One Sample Test Example

- ▶ Consider the scenario where this collection of cars *is* in fact sampled from a population having a mean average mileage equal to 10km/L. (i.e. our null hypothesis is correct).

- ▶ Based on the randomness of sampling, it would be unrealistic to expect that our sample would produce an $\overline{x}$ exactly equal to 10km/L.

- ▶ Therefore we should expect some plausible wiggle room around our hypothesized value of 10 which we would deem close enough.

One Sample Test Example

- On the contrary, once we pass a certain threshold, we could determine that our sample is inconsistent with our hypothesis.

- That threshold should of course depend on the problem at hand.

- For example, we would explect the threshold for the black curve on this slide should would be more strict than the thresold for the green curve.

One Sample Test: Hypotheses Statements

- In the car mileage example, we are testing:

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

that is to say, our hypothesized value is $\mu_0 = 10$.

- Analogous to the courtroom, we believe H_0 is true until evidence (in the form of data) suggests otherwise.

- To determine our threshold which determines how far our sample mean needs to depart from 10 before we no longer believe H_0 is true, we need a **test statistic**.
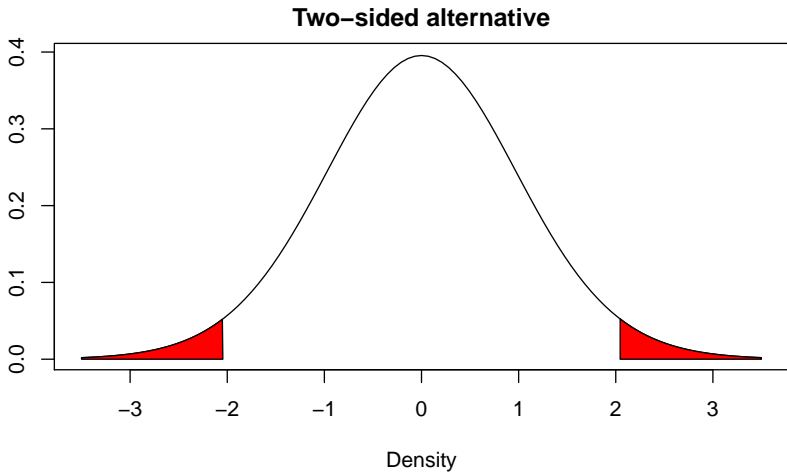
One Sample Test: Calculate Test Statistic

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

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

$\overline{x}$ is a sample mean, s is sample standard deviation, n is sample size, μ_0 is hypothesized mean value

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

The distribution of (1) looks like this:

Two–sided alternative

- If the hypothesis is true, there is only a 5% chance that our test statistic falls in the area in red.

- This may seem like a reasonable way to determine our threshold for believing $H_0 = 10$.

- To put numbers to this example, if $\overline{x}$ is less than -9.6446667 or bigger than 10.3553333 we will no longer believe the null hypothesis that the average gas mileage for this population of cars is equal to 10km/L.

- Indeed this line of thinking is exactly what we follow when forming conclusions for a *two-sided* `t.test`.

- A similar argument can be made for one-sided tests:
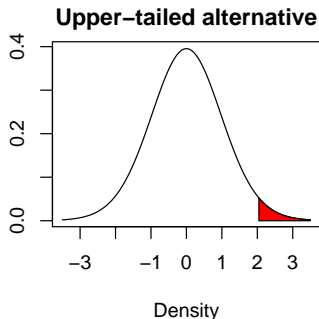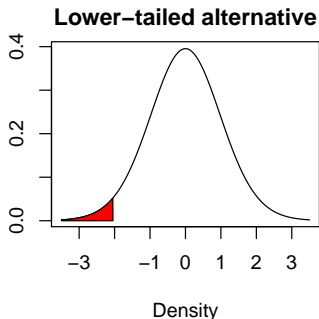
- This area in red is sometimes referred to as the **rejection region**.

Hypothesis testing in R

- Now that we have an understanding of what these tests are doing, lets see how we can code them up in R.

- The general syntax for performing t.test in R is:

    ```
    t.test(x=mydata,alternative="two.sided",mu = μ₀
                    ,conf.level=0.95,...,)
    ```

- In R the default `alterative` is `"two.sided"` with the options `"less"`, or `"greater"`.

- As usually, to see the help file on this function, we type `?t.test`.

Hypothesis testing in R

- The `conf.level` = 1 - α, where α denotes our significance level.

- The **significance level** determines how much of the probability we allow in the rejection region.

- Typically $\alpha = 5\%$ so that the `conf.level` = 95% (or 0.95)

- A confidence level of 0.95 is the default setting in R, hence we will usually not specify it in the code to follow.

One Sample Test: Cars Example

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

$$H_0 : \mu = 10 \qquad \text{vs} \qquad H_1 : \mu \neq 10$$

```r
car_data <- read.csv("car_data.csv") # read in the data
t.test(x=car_data$km.L,               # sample mileage
        alternative=c("two.sided"),   # two-side H_A
        mu=10) # set the hypothesized value equal to 10.

# since a two-sided test is the default, we could have typed:
t.test(x=car_data$km.L, mu=10)
```

R Output

```
##
##  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
```

Decision and Conclusion (using P-value)

- In order to make a conclusion based on this output, we usually look at a p-value.

- The **p-value**, is the probability of sampling data more extreme than what we observed when the null hypothesis H_0 is true.

- Hence the smaller this value is, the more unlikely our null hypothesis is true.

- **Question**: How small is *too small*?

    - **Answer**: Anything less that the significance level α, we deem too small.

- **Why?** This *p*-value has a one-to-one relationship with the rejection region method described earlier.

- To be more specific, if our sample surpasses our threshold, the *p*-value will necessarily be less than α.

Decision and Conclusion (using *P*-value)

Assuming our α/significance level is 5%. . .

If $p - \text{value} > 0.05$, the probability of seeing a sample mean more extreme than $\overline{x}$ is not that unlikely.

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 $p-\text{value} < 0.05$, the probability of seeing a sample mean more exteme than $\overline{x}$ is unlikely.

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).

Under the hood calculations

- Returning to our car example, recall that we are testing:
  - $H_0 : \mu = 10$ vs. $H_1 : \mu \neq 10$.
- Intuitively, as our sample mean $\overline{x}$ gets father and rather away from 10, this would indicate stronger and stronger evidence against the null hypothesis.
- We can see calculate our test statistic:

$$\frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{10.3553333 - 10}{\frac{1.210354}{\sqrt{30}}} = 1.6079931$$

and find the corresponding p-value $= 0.11867$.

Two sided test

For the visual learners, we can view our *p*-value as follows:

Upper–tailed alternative

Density
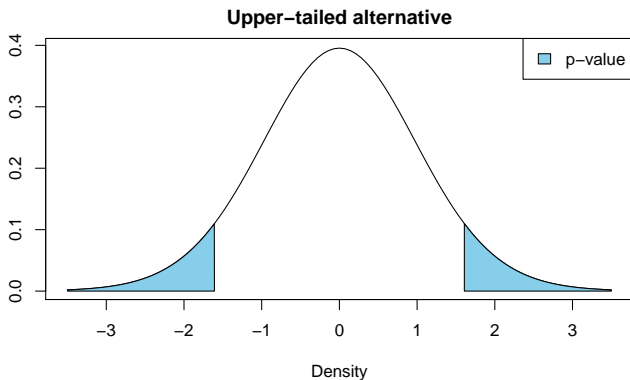
Alternative ways of thinking

► Since this blue area on the previous slide makes up more than 5%, we fail to reject the null hypothesis.

► Since the probability of observing a sample mean more extreme than the one we saw (i.e > 10.355 or < -10.355) is quite likely (approximate 11.9%), we fail to reject the null hypothesis.

Decision and Conclusions (using *P*-value)

- While it is important to understand the "under-the-hood calculations" (see STAT 230 for more) all of our decisions for this course may be based on the R output.

- Since the *p*-value of 0.11867 is greater than our significance level of 0.05, we fail to reject the null hypothesis.

R output:

```
                        R code

> t.test(x=car_data$km.L, mu=10)


            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
```

Decision and Conclusions (using *P*-value)

▶ We usually like to state our conclusions in terms of the problem at hand,. . .

▶ Hence, there is insufficient evidence to suggest that the mean mileage differs from 10 km/L.

▶ N.B: We never claim that either the null or alternative hypothesis is true. We can only "reject" or "fail to reject" the null hypothesis.

- Before moving on to the two-sample problem, it is worth discussing the **confidence interval** which is also produced as output for a t-test.

- This interval gives us a range of plausible values for μ (regardless of our hypothesized value μ_0)

General Form of Confidence Interval (CI):

The general form of a confidence interval looks like this:

$$\text{point estimate} \pm \text{margin of error}$$

$$(\hat{\mu} - m.e., \ \hat{\mu} + m.e.)$$

Interpretation: Assuming a confidence level of 95%, we are 95% confident that the interval will contain the true value of the parameter.

Decision and Conclusions (using CI)

```
                    ── R code ──
> t.test(x=car_data$km.L, mu=10)

            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
```

When the input argument `conf.level`=0.95 = (1-α) in the `t.test` function, a 95% confidence interval is produce.

Other not-so-typical alternatives might be:

- ▶ setting `conf.level`=0.90, and obtaining a 90% confidence interval.

- ▶ setting `conf.level`=0.98, and obtaining a 98% confidence interval.

- ▶ setting `conf.level`=0.99, and obtaining a 99% confidence interval.

- We can actually make a conclusion (reject or fail to reject) based on the confidence interval instead of looking at the *p*-value.

- For the car example, we are 95% confident that the true mean mileage of the car lies within the bounds $[9.903, 10.807]$.

- Since 10 is within those bounds, the hypothesis that $\mu = 10$ would appear reasonable; hence we fail to reject H_0.

- If, for example, our CI was $[15.3, 16.4]$ instead, this hypothesis is not within reason and we would reject the null hypothesis.

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 may be assumed to come from distributions with 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:

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

- An equivalent way of saying this is:

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

- More generically, we could test:

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

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.

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

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:

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

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

Hypothesis testing in R

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

  ```
  t.test(x=A, y=B,alternative="two.sided",mu = d_0
         ,conf.level=0.95, data=mydata, ...)
  ```

- Alternatively, we could specify a *formula*

  ```
  t.test(formula=A∼B,alternative="two.sided",mu = d_0
         ,conf.level=0.95, data=mydata, ...)
  ```

- Where `A` and `B` contain the samples from group A and B, respectively.

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:

```
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:

```
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.[1]

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:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \textit{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

  whereas the formula notation is testing:

$$H_0 : \mu_2 - \mu_1 = 0 \quad \textit{vs.} \quad H_1 : \mu_2 - \mu_1 \neq 0$$

- 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:
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
```

- Based on the *p*-value provided in the summary = $7.2691124 \times 10^{-31}$, there is very strong evidence to suggest that the average temperature between active and non-active beavers are different from one another.

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

- Note that we still have the relationship between the two-sided test and the corrresponding confidence interval.

- The CI is highlighted in pink in the following slide ...

Two Sample Unpaired Example

```
┌─────────────────── Using CI-value ───────────────────┐
│ > t.text(temp-activ, data=beaver2, alternative=c("two.sided")
│ paired=FALSE)
│         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
│ 95 percent condifence interval:
│   -0.8927106  -0.7197342
│ sample estimates:
│ mean in group 0 mean in group 1
│        37.09684         37.90306
└──────────────────────────────────────────────────────┘
```

- The two sample case tests a DIFFERENCE between the groups ($\mu_d = \mu_2 - \mu_1 \neq 0$).

- The CI stated on the previous slide is the CI for the difference, $\mu_d = \mu_1 - \mu_2$

- We reject the null hypothesis because 0 is not contained in the interval.

- If 0 was contained we would fail to reject the null hypothesis.

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. Apply something to test subjects to see if there is an effect

3. Car example: 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. In other words, test to see if the mean of the 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("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

```
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

```
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
```

Two Sample Paired Test Example

Decision and Conclusion

- Alternatively we could have arrived at this conclusion through confidence intervals.

- The two-sample t-tests provides a 95% CI for $\mu_d = [-0.554, 0.498]$.

- Since the hypothesized difference of $d_0 = 0$ lies within this interval, we fail to reject the null hypothesis.

Question

Which (if any) of the following are true?

1. Paired and unpaired t-tests are the same thing. ✗

2. Confidence intervals can be of any level of confidence (not just 95%). ✓

3. Confidence intervals can be used to make a conclusion about a hypothesis test. ✓

4. Confidence intervals can be used to prove that the null hypothesis is false. ✗

Question

Which (if any) of the following are true?

1. Unpaired t-tests test the difference between two means μ_1 and μ_2. ✔

2. Paired t-tests can be used to compare the difference between two measurements on the same subject. ✔

3. In both the paired and unpaired two sample cases, a confidence interval containing 0 would result in a decision of: fail to reject the null hypothesis. ✗

4. In the one sample t-test, a confidence interval containing 0 would result in a decision of: fail to reject the null hypothesis. ✗

Question

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

1. Is the average final grade for Data 301 greater than 70%? *one sample t-test*

2. Does a student's mark improve after studying (measurements taken on same student)? *two sample paired*

3. Has the average student height increased since 1990? *two sample unpaired (two distinct student populations)*

4. Does radiation reduce the size of tumours when used to treat patients? *two sample paired (same patient), although could argue against control/experiment groups in which case two sample unpaired*

Question

Using the car data, test the hypothesis that the mean distance traveled at each fill up is less than 450km.

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:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \qquad (2)$$

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 attempt 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 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
```

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 STAT 230 and DATA 311.