LESSON: Hypothesis testing

FOCUS QUESTION: How can I tell whether the test group is different from the control group?

In this lesson you will:

- Formulate and test a hypothesis regarding population mean.
- Apply the one sample t-test to assess the true mean.
- Apply the two sample t-test to assess whether two samples are likely to come from populations with the same mean.
- Use p-values and confidence intervals.

Contents

- **SUGGESTED READING:** Wikipedia has a discussion of hypothesis testing
- DATA FOR THIS LESSON
- SETUP FOR LESSON
- EXAMPLE 1: Load the consolidated sleep diary data
- EXAMPLE 2: Does subject 1 sleep 8 hours on average?
- EXAMPLE 3: Do students in section 2 sleep 8 hours on average?
- EXAMPLE 4: Do the students in sections 2 and 3 sleep a different amount?
- EXAMPLE 5: Do section 2 and 3 students sleep differently at the 0.01 significance level?
- EXAMPLE 6: Do section 2 students sleep more than section 3 students?
- EXAMPLE 7: Do section 2 students sleep more than section 3 students (fewer assumptions)?
- SUMMARY OF SYNTAX

SUGGESTED READING: Wikipedia has a discussion of hypothesis testing

SUGGESTED READING: Wikipedia also has a discussion of the concept of the null hypothesis which is somewhat readable. The discussion can be found at <http://en.wikipedia.org/wiki/Null_hypothesis>.

SUGGESTED READING: Wikipedia discusses the meaning of the p-value and the frequent misunderstandings in interpreting it. The discussion can be found at <http://en.wikipedia.org/wiki/Pvalue>.

DATA FOR THIS LESSON

File	Description

`diaries.mat` (found on Learn)	▪ The data set contains sleep diary data for a cohort in MATLAB variables.
	▪ The arrays have a column for each person.
	▪ The vectors have an element for each person.
	▪ The values in column *n* correspond to the same person as the value in position *n* of each vector.
	▪ The file contains the following variables:
	  ▪ `bedTimes` - array of bed times in decimal-date format.
	  ▪ `dayCaffeine` - array of daytime caffeine indicators.
	  ▪ `gender` - vector of male/female gender designators.
	  ▪ `nightCaffeine` - array of evening caffeine indicators.
	  ▪ `section` - vector of section indicators. The possible section numbers are 0, 1, 2, and 3. Section 0 contains only a single instructor. The remaining values correspond to course section numbers.
	  ▪ `toSleepMinutes` - an array of number of minutes to fall asleep.
	  ▪ `useAlarm` - array of alarm use indicators.
	  ▪ `wakeTimes` - array of wakeup times in decimal-date format.
	▪ The data was originally gathered by students taking CS 1173 in the fall 2009 semester and anonymized and randomized to be unidentifiable.
	▪ The first column of each array represents the instructor's values, the rest of the columns represent individual students.
	▪ Diaries were recorded for 21 days (from September 23, 2009 to October 13, 2009).

SETUP FOR LESSON

- Create a `HypothesisTesting` directory on your `V:` drive and make it your current directory.
- Download the `diaries.mat` data file from Blackboard and save it to your `HypothesisTesting` directory.
- Create a `HypothesisTestingLesson.m` script file in your `HypothesisTesting` directory. Enter each of the examples in a new cell in this script.

EXAMPLE 1: Load the consolidated sleep diary data

Create a new cell in which you type and execute:

```
load diaries.mat;  % Load the sleep diaries

sleepHours = (wakeTimes - bedTimes)*24;  % Calculate hours of sleep
```

You should see 9 variables in the Workspace Browser. We will be interested in the following variables:

- `gender` - vector containing gender of the individual subjects
- `section` - vector containing sections numbers of the individual subjects
- `sleepHours` - an array number of hours of sleep of individuals

EXAMPLE 2: Does subject 1 sleep 8 hours on average?

Create a new cell in which you type and execute:

```
[h1, p1, c1] = ttest(sleepHours(:, 1), 8);

fprintf(['Does subject 1 sleep 8 hours on average?\n\t' ...

    'h = %g, p = %g, ci = [%g, %g]\n'], h1, p1, c1);
```

You should see 3 variables in the Workspace Browser:

- `c1` - confidence interval for the difference of the two population means
- `h1` - a value indicating whether to reject (1) or not reject (0) the null hypothesis
- `p1` - (the p-value) gives the probability that such a sample could be picked by chance if the null hypothesis were really true

You should also see the following output:

```
Does subject 1 sleep 8 hours on average?

        h = 0, p = 0.113092, ci = [7.21123, 8.09036]
```

EXERCISE 1: What is the null hypothesis for EXAMPLE 2?

EXERCISE 2: Is the null hypothesis rejected for EXAMPLE 2?

EXAMPLE 3: Do students in section 2 sleep 8 hours on average?

Create a new cell in which you type and execute:

```
sleepHoursSec2 = sleepHours(:, section == 2);

[h2, p2, c2] = ttest(sleepHoursSec2(:), 8);

fprintf(['Do section 2 students sleep 8 hours on average?\n\t' ...
```

```
        'h = %g, p = %g, ci = [%g, %g]\n'], h2, p2, c2);
```

You should see 4 variables in the Workspace Browser:

- c2 - confidence interval for the difference of the two population means
- h2 - a value indicating whether to reject (1) or not reject (0) the null hypothesis
- p2 - (the p-value) gives the probability that such a sample could be picked by chance if the null hypothesis were really true
- sleepHoursSec2 - hours of sleep for students in section 2

You should also see the following output:

```
Do section 2 students sleep 8 hours on average?

        h = 1, p = 1.91292e-06, ci = [8.30087, 8.71824]
```

EXERCISE 3: What is the null hypothesis for EXAMPLE 3?

EXERCISE 4: What is the alternative hypothesis for EXAMPLE 3?

EXERCISE 5: Is the null hypothesis rejected in favor of the alternative hypothesis for EXAMPLE 3?

EXAMPLE 4: Do the students in sections 2 and 3 sleep a different amount?

Create a new cell in which you type and execute:

```
    sleepHoursSec3 = sleepHours(:, section == 3);

    [h3, p3, c3] = ttest2(sleepHoursSec2(:), sleepHoursSec3(:));

    fprintf(['Do students in sections 2 and 3 get different amounts of sleep on
average?\n\t' ...

        'h = %g, p = %g, ci = [%g, %g]\n'], h3, p3, c3);
```

You should see the following variables in your Workspace Browser:

- c3 - confidence interval for the difference of the two population means
- h3 - a value indicating whether to reject (1) or not reject (0) the null hypothesis
- p3 - (the p-value) gives the probability that such a sample could be picked by chance if the null hypothesis were really true

- `sleepHoursSec3` - hours of sleep for students in section 3

```
Do students in sections 2 and 3 get different amounts of sleep on average?

        h = 1, p = 0.0035873, ci = [0.127686, 0.652328]
```

EXAMPLE 5: Do section 2 and 3 students sleep differently at the 0.01 significance level?

Create a new cell in which you type and execute:

```
    [h4, p4, c4] = ttest2(sleepHoursSec2(:), sleepHoursSec3(:), 0.01);

    fprintf(['Do students in sections 2 and 3 sleep differently at the 0.01
significance level?\n\t' ...

        'h = %g, p = %g, ci = [%g, %g]\n'], h4, p4, c4);
```

You should see the following variables in your Workspace Browser:

- c3 - confidence interval for the difference of the two population means
- h3 - a value indicating whether to reject (1) or not reject (0) the null hypothesis
- p3 - (the p-value) gives the probability that such a sample could be picked by chance if the null hypothesis were really true

```
Do students in sections 2 and 3 sleep differently at the 0.01 significance level?

        h = 1, p = 0.0035873, ci = [0.0451417, 0.734872]
```

EXAMPLE 6: Do section 2 students sleep more than section 3 students?

Create a new cell in which you type and execute:

```
   [h5, p5, c5] = ttest2(sleepHoursSec2(:), sleepHoursSec3(:), 0.05, 'right');

    fprintf(['Do sections 2 students get more sleep than section 3 students?\n\t' ...

        'h = %g, p = %g, ci = [%g, %g]\n'], h5, p5, c5);
```

You should see the following 3 variables in your Workspace Browser:

- c5 - confidence interval for the difference of the two population means
- h5 - a value indicating whether to reject (1) or not reject (0) the null hypothesis

- p5 - (the p-value) gives the probability that such a sample could be picked by chance if the null hypothesis were really true

```
Do sections 2 students get more sleep than section 3 students?

        h = 1, p = 0.00179365, ci = [0.169891, Inf]
```

EXAMPLE 7: Do section 2 students sleep more than section 3 students (fewer assumptions)?

Create a new cell in which you type and execute:

```
[h6, p6, c6] = ttest2(sleepHoursSec2(:), sleepHoursSec3(:), 0.05, 'right', 'unequal');

fprintf(['Do sections 2 students get more sleep than section 3 students?\n\t' ...

    'h = %g, p = %g, ci = [%g, %g]\n'], h6, p6, c6);
```

You should see the following 3 variables in your Workspace Browser:

- c6 - confidence interval for the difference of the two population means
- h6 - a value indicating whether to reject (1) or not reject (0) the null hypothesis
- p6 - (the p-value) gives the probability that such a sample could be picked by chance if the null hypothesis were really true

```
Do sections 2 students get more sleep than section 3 students?

        h = 1, p = 0.00198515, ci = [0.167467, Inf]
```

SUMMARY OF SYNTAX

MATLAB syntax	Description
h = ttest(X, m)	Perform a one-sample student's t-test to determine whether the true mean of the population represented by the sample in the vector X could have a value different than m. The significance level for the test is 0.05. If h is 1, then it is likely that the mean of the population represented by the sample X is different from m. If h is 0, then you don't have enough evidence to conclude that the mean is different from m.

MATLAB syntax	Description
	The `ttest` assumes that `X` is a random sample drawn from a normally distributed population. If `X` is an array, `ttest` works along the first non-singleton dimension. **Note: Do NOT take the mean of `X` before applying `ttest`.**
`[h, p, ci] = ttest(X, m)`	Perform a one-sample student's t-test to determine whether the true mean of the population represented by the sample in the vector `X` is `m`. The variable `p` represents a *p-value*, indicating how likely it is to observe the test statistic if the population mean were actually equal to `m`. The variable `ci` holds the 95% confidence interval for the true mean.
`[h, p, ci] = ttest(X, m, alpha)`	Perform a one-sample student's t-test at significance level `alpha` to determine whether the true mean of the population represented by the sample in the vector `X` is is different from `m`. The variable `p` represents a *p-value*, indicating how likely it is to observe the test statistic if the population mean were actually equal to `m`. The variable `ci` holds the 100*[1 - `alpha`]% confidence interval for the true mean.
`[h, p, ci] = ttest(X, m, alpha, 'left')`	Perform a one-sided one-sample student's t-test at significance level `alpha` to determine whether the true mean of the population represented by the sample in the vector `X` is different from `m`. If `h` is 1, then it is likely that the mean of the population represented by the sample `X` is less than `m`. The variable `p` represents a *p-value*, indicating how likely it is to observe the test statistic if the population mean were actually equal to `m`. The variable `ci` holds the 100*[1 - `alpha`]% confidence interval for the true mean.
`[h, p, ci] = ttest(X, m, alpha, 'right')`	Perform a one-sided one-sample student's t-test at significance level `alpha` to determine whether the true mean of the population represented by the sample in the vector `X` is different from `m`. % If `h` is 1, then it is likely that the mean of the population represented by the sample `X` is greater

MATLAB syntax	Description
	than `m`. The variable `p` represents a *p-value*, indicating how likely it is to observe the test statistic if the population mean were actually equal to `m`. The variable `ci` holds the 100*[1 - `alpha`]% confidence interval for the true mean.
`h = ttest2(X, Y)`	Perform a two-sample student's t-test to determine whether the true means of the populations represented by the samples X and Y are different. If `h` is 1, then it is likely that the means of the respective populations represented by samples X and Y are different. If `h` is 0, then you don't have enough evidence to conclude that the means are different. The significance level for the test is 0.05. The `ttest2` assumes that X and Y are random samples drawn from a normally distributed populations. If X is an array `ttest2` works along the first non-singleton dimension. In this case Y must be the same size as X except along the first non-singleton dimension. **Note: Do NOT take the mean of X or of Y before applying `ttest2`.**
`[h, p, ci] = ttest2(X, Y)`	Perform a two-sample student's t-test to determine whether the true means of the populations represented by the samples X and Y are different. The variable `p` represents a *p-value*, indicating how likely it is to observe the test statistic if the population means were actually equal. The variable `ci` is the 95% confidence interval for the difference of the two population means.
`[h, p, ci] = ttest2(X, Y, alpha)`	Perform a two-sample student's t-test at significance level `alpha` to determine whether the true means of the populations represented by the samples X and Y are different. The variable `p` represents a *p-value*, indicating how likely it is to observe the test statistic if the population meand were actually equal. The variable `ci` holds the 100*[1 - `alpha`]% confidence interval for difference of the true population means.

MATLAB syntax	Description
`[h, p, ci] = ttest2(X, Y, alpha, 'left')`	Perform a one-sided two-sample student's t-test at significance level `alpha` to determine whether the true mean of the population represented by the sample `X` is less than the true mean of the population represented by the sample `Y`. If `h` is 1, then it is likely that the mean of the population represented by the sample `X` is less than the population mean represented by the sample `Y`. The variable `p` represents a *p-value*, indicating how likely it is to observe the test statistic if the population mean corresponding to `X` were actually greater than or equal to the population mean corresponding to `Y`. The variable `ci` holds the 100*[1 - `alpha`]% confidence interval for the difference of the two populations means.
`[h, p, ci] = ttest2(X, Y, alpha, 'right')`	Perform a one-sided two-sample student's t-test at significance level `alpha` to determine whether the true mean of the population represented by the sample `X` is greater than the true mean of the population represented by the sample `Y`. If `h` is 1, then it is likely that the mean of the population represented by the sample `X` is less than the population mean represented by the sample `Y`. The variable `p` represents a *p-value*, indicating how likely it is to observe the test statistic if the population mean corresponding to `X` were actually less than or equal to the population mean corresponding to `Y`. The variable `ci` holds the 100*[1 - `alpha`]% confidence interval for the difference of the two populations means.

This lesson was written by Kay A. Robbins of the University of Texas at San Antonio and last modified by Dawn Roberson on 3 Nov-2013. Please contact krobbins@cs.utsa.edu with comments or suggestions. The photo is of Sir Ronald Fisher, founder of modern statistics and namesake of the Fisher Iris dataset. (See http://en.wikipedia.org/wiki/File:R._A._Fischer.jpg.

Published with MATLAB® 7.14