M Gmail   **Andrew Ho <kironide@gmail.com>**

Week 2 Day 4 (Morning)
3 messages

Jonah Sinick <jsinick@gmail.com>   Thu, Feb 25, 2016 at 2:42 AM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft
<clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani
<jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew
Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Sam Eisenstat
<sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

Today we're going to look at comparison of means between groups.

- Read the Introduction and Test 1 through Test 10 of 100 Statistical Tests by Gopal Kanji, writing a
  function in R to perform each test, with descriptive variable names, in order to make sure that you
  understand the statements of the tests. Don't worry about memorizing them.

- Read Calculating p-values (in R).

- Skim Always use Welch's t-test instead of Student's t-test from "The 20% Statistician."

- Check out the documentation for pairwise.t.test in R https://stat.ethz.ch/R-manual/R-
  patched/library/stats/html/pairwise.t.test.html. Read about the Bonferroni Correction https://en.
  wikipedia.org/wiki/Bonferroni_correction (pairwise.t.test uses a modified version).

- Load the dataset "singer" from the "lattice" package. Compute the means by voice part. Apply
  pairwise.t.test to the dataset to determine which differences in height by voice part are statistically
  significant.

- If you haven't finished *Advanced R,* Chapters 1–6, 10 and 11, continue working through them (including the
  exercises).

- Take a look at LWSurveyExploration.zip (as attached). Explore groupings of the numeric data by
  category, using t.test and pairwise.t.test as appropriate. Write up your most interesting finding in a .Rmd
  file.

📁 **LWSurveyExploration.zip**
488K

Jonah Sinick <jsinick@gmail.com>   Thu, Feb 25, 2016 at 10:13 AM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft
<clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani
<jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew
Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Sam Eisenstat
<sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

The case study at the end of the tutorial that http://www.cyclismo.org/tutorial/R/pValues.html is from is also
good for a real world example using t-tests and calculating p-values: http://www.cyclismo.org/
tutorial/R/cholesterol.html
[Quoted text hidden]

Jonah Sinick <jsinick@gmail.com>   Thu, Feb 25, 2016 at 3:20 PM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft

<clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

I've attached the presentation.

A correction thanks to Ali: The standard deviation of the sampling means is standardDeviationOfSample/ sqrt(sampleSize - 1): I was missing the -1.

You can look at the movielens dataset: http://grouplens.org/datasets/movielens/ (I used the one with 1,000,000 ratings).

Main question: Which movies are robustly higher rated by women?

Write up your analysis as a .Rmd folder.

Feel free to explore genre information.

Watch out for the multiple comparisons problem

If you'd like to do something else with the dataset, you can check out recommenderlab: A Framework for Developing and Testing Recommendation Algorithms ;-).
[Quoted text hidden]

---

2 attachments

movieGenderDifferences.html
1057K

movieGenderDifferences.Rmd
3K