Resampling, bootstrap and permutation testing

Jeff LeDue

7 April 2017

First… Survey results.

- Thanks to everyone who took the time to fill out the survey, Response rate: 44%.
-
- Q1: >50% have attended >7 sessions.
- Q2: 76% Hit or miss, 24% learning a lot
- Q3: 79% useful to have materials during or after the sessions.

Survey con't

- Q4: Continue in the summer: 73%
- This means we need volunteers!
-
- Topic most often suggested: stats.
- People want stats. Particularly a stats primer.
- Mentioned in 1/3 of responses to this question. (what stats test mean)
-
- Part of the reason I am doing today's databinge on resampling.
-

Full list of Suggested Topics

- (posted on slack):
- Camera technology, PCA, clustering, micromanager, CRISPR/cas9, more computer coding, more matlab, biological/genetic/molecular-based tools, Genetics, imaging, two-photon, Physics and Electrical Engineering of Ephys, basics of cloning, Data scraping, Gene ontology resources, Allen brain institute resources, LaTeX, Google Scholar and other citation alert systems, Bash/Shell scripting, graphic design software, ICA, strobing, blood volume signal removing, three color imaging, automatic optogenetics, How to use publicly available "big data" sets (RNA-seq, etc.), Functional/live brain imaging/brain slice imaging/in vivo imaging and analysis, analyzing behaviour data

Other points of interest

- A databinge award. Introducing the Orange Prize. ?
-
- Story arcs. We have only had one so far which was Mouse Genetics – GCAMPs – AAV, which was a little out of order.
  - More detail, spread out a topic over ~3 weeks + 1 week if anyone wants help to apply any of the ideas discussed to real data from their projects
-
- I suggest the current one on Stats/Analysis methods ?
  - Wk1: Resampling, bootstrap and permutation testing
  - Wk2: Statistical tests primer?
  - Wk3: PCA/ICA? Or GLMEM?
  - Wk4: Data anyone?

Outline

- Why?
-
- What is resampling?
- The bootstrap distribution. What can we get from it?
  - Bootstrap standard error
  - Bootstrap CI's
- Hypothesis testing with resampling.
  - permutation testing

Why did I learn about this?

- For a project in Tim's lab with a former graduate student.
- We wanted to test whether or not there were differences in connectivity between different brain regions.
- Our data did not look like a Gaussian :(
-
- Tried permutation testing since we'd seen it used in other papers. It's distribution free!
- Found Ch18! Very readable intro to these concepts.

Why should we all learn this?

- Randomization is a good way to explore your data.
- Nate mentioned this idea in his talk: for sanity checking.
  - What that aspect is exactly and whether or not it is a good or bad idea to randomize on it is a subject for debate in the lab.
- If you do the randomization a lot of times, it gives you a sense of how “unusual” your observed data is.
- These randomization methods have few assumptions, are general, and a good way to get an intuitive feel for your data (& statistics in general)
-
- Also:
- Your reviewers/committee/supervisor could/should/will ask you to make bootstrap confidence intervals.
- It happens.
- You will say, “No problem”.

So what is a resample?

- This refers to the generation of a new sample from your existing data.
- You do this by sampling with replacement. In practice you need a computer for this since you ask it to generate random numbers which determine which of your data ends up in your resample.
- You generate 100's or 1000's of these resamples.

The bootstrap idea

- Page 18-8
- “The original sample represents the population from which it was drawn. So resamples from this sample represent what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, represents the sampling distribution of the statistic, based on many samples.”

Ok, so why do we want these resamples?

- The resamples are useful because you can calculate any sample statistic (eg sample mean) for each of them.
- This gives you a distribution of possible sample means. This is called the bootstrap distribution.
- It's useful as an estimate for the sampling distribution.
- The sampling distribution is what tells you about the most likely value and the range of values for any statistic
- We are used to getting these from theory in the form of tables, computers.
- But here we have an estimate for it numerically and we can make use of it to estimate:
  - standard errors
  - confidence intervals.
-
- Jump over to matlab and do some of this stuff.

On the F103 PC's...

- The materials will be in E:\Databinge_Resampling
- Open Matlab and navigate to this folder.
- Open “jeffs_script_working_copy.m”
- This is a matlab file with lots of comments describing what we will do. Kind of fill in the blanks with the code.
- There is an already filled in version called “jeffs_script.m”
- We will write a couple of functions. Working copies of these are in the subfolder “working functions”

Bootstrap standard error & confidence intervals

- As we just say in matlab:
- We are using the bootstrap distribution as a stand in for the sampling distribution its mean is an estimator for the expected value of that statistic.
- So continuing this, the width (standard deviation) of the bootstrap distribution is the standard error of the statistic.
- And we can calculate confidence intervals in a number of ways:
  - Sampling distribution looking Normals: studentized
  - Samp dist. Not Normal? Try:
    - Percentiles or Bias Corrected and Accelerated CI.

An uneasy feeling...

- Page 18-11: Why does bootstrapping work?
- “It might seem that the bootstrap creates data out of nothing. This seems suspicious. But we are not using the resampled observations as if they were real data—the bootstrap is not a substitute for gathering more data to improve accuracy.”
- Page 18-12 “Using the data twice—once to estimate the population mean, and again to estimate the variation in the sample mean—is perfectly legitimate.”

Permutation testing

- Extend the idea of resampling to hypothesis testing.
- **Step1:** Determine a statistic that addresses what you want to test (ie difference of the means of two samples). Calculate the observed value.
- **Step2:** Resample in a way that is consistent with the Null hypothesis. (ie if your Null hypothesis is true then there won't be any difference if you draw your resamples from all the data, regardless of the original sample group)
- **Step 3:** Make a distribution of the statistic using your resamples. See where you observed value falls on this distribution. This tells you how about rare your particular observation is and is a direct way to estimate a p-value.
-
- Jump back to Matlab to try this.