Hypothesis Testing and Central Limit Theorem
Exercises in R

January 9, 2018

1 Demonstration of Central Limit Theorem

In this exercise we would like you to explore some of the implications of the central limit theorem (CLT). In the first portion you will write a function in R that demonstrates how a normal distribution can be used to approximate a discrete distribution. In the second part, you will investigate the implications of the CLT's assumptions. For this section (and all others) please include your code in an appendix.

1.1 Approximating Binomial

We would like to write a function in R that simulates tossing n fair coins and counting the number of "heads".[1] Each set of tosses of n coins constitutes a sample; and we are interested in the number of "heads" in each sample.

We would like our function to take 2 arguments, `n`, and `n_sims`, where `n` is the number of coins, and `n_sims` is the number of samples. We would also like the function to return a vector of length `n_sims` where each element is the number of "heads" in the corresponding sample.

Once we have our function, we will use it to run some simulations with different values of n. In particular, we would like to run simulations with $n = 5$, $n = 10$, $n = 30$, and $n = 50$. For now, we will set `n_sims` to 1000.

[1] The base R language has an exceptionally rich set of functions for simulating data from many discrete and continuous distribution. For example, the `rnorm()` function can be used to take random draws from the normal distribution.

We would also like to plot the results of these simulations. In particular, we want to generate a histogram showing the density of our simulations. Additionally, we would like to overlay a normal distribution with the appropriate mean and standard distribution on top of these four histograms.[2]

We are also interested in the behavior of the distribution when we change the number of simulation iterations to be larger or smaller than 1000. What happens when `n_sim = 100` and `n_sim = 10000`?

To summarize, for this section, please complete the following:

1. Write a function to simulate tossing n fair coins and counting the occurrence of "heads"
2. Run that function for $n = 5$, $n = 10$, $n = 30$, and $n = 50$, and `n_sims` equal to 1000
3. Plot the density of the resulting counts and overlay the normal distribution with appropriate mean and standard deviation
4. Describe briefly what you notice in the plots
5. Run your simulation function again for $n = 5$, $n = 10$, $n = 30$, and $n = 50$, but this time use `n_sims` of 100 and `n_sims` of 10000
    - Plot histograms of the counts and describe what you notice
    - How do the different values of `n_sims` impact the density
    - There is no need to include these additional plots in your write up, just describe them

1.2 Sampling from a Cauchy

Suppose that now we are interested in the behavior of sample means from a different distribution. In particular, in this exercise we will write a function—similar to the one above—that computes sample means from random draws taken from a Cauchy distribution with location 0 and scale 1.

In particular, we want to generate samples of size $n = 100$ and take the mean of these samples over `n_sims` equal to 1000 iterations. Plot a histogram of the resulting means. Briefly describe what you notice. Is this behavior similar to what we noticed previously? What might be a potential explanation for the behavior?

[2] We encourage exploration of the plotting features in R. The default plots are not particularly aesthetically pleasing, but with a small amount of effort, R can produce very beautiful plots.

2 Diabetic Patient Admissions

When a patient is admitted to a hospital, their admission is generally categorized according to the severity or urgency of their symptoms (e.g., "*Elective*", "*Emergency*", etc.). We are interested in the different types of hospital admissions for diabetic patients. We will use the data in the file `diabetes_data_clean.csv` to answer several questions related to hospital admissions.[3] Use the appropriate statistical tests, plots, or tables to address the questions below.

2.1 Admissions by Gender

Using the `diabetes_data_clean.csv` data, does admission type seem to differ by gender? If so, in what way? If applicable, determine the appropriate statistical test and describe the distributions for females and males.

Suppose now that we are only concerned with hospital admissions in the `Elective` and `Emergency` categories. Does it seems that males and females differ with respect to these?

What happens if we only consider those hospital admissions that were the result of a physician referral? In particular, if we consider only those hospital admissions that were from physician referrals, do we find that females and males differ in whether their hospital admissions were `Elective` or `Emergency`?

3 Diabetic Medications

The data in `diabetes_data_clean.csv` has a variable with information about the number of medications for the patient at each visit. We will be interested in whether the nature of the patient's admission impacts the number of medications. We will also be interested in the change in number of medications over time.

We will be working with some continuous data in this section, so please include plots with your responses.

[3] The data were obtained from the University of California Irvine's Machine Learning repository. The raw data can be obtained here: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008. Note that the version we are using in `diabetes_data_clean.csv` has been somewhat cleaned up. The data contain over 100,000 hospital admission records for diabetic patients from across 130 United States hospitals from 1999 to 2008

3.1 Medications by Admissions

We are interested whether patients differ in the number of medications based on whether their admission type was `Elective` or `Emergency`. Using the appropriate statistical test, determine whether or not patients coming in electively or in the case of an emergency tend to receive the same number of medications. Note that some filtering of data will be required.

What are the assumptions of the statistical test we use here? Are we confident that these assumptions are satisfied?

3.2 Medications and Repeat Visits

Note that the data in `diabetes_data_clean.csv` has patients who had repeat visits to the hospital. We know this to be true because `patient_nbr` is the unique patient identifier, and there are instances of recurrence.

We would like to see whether the number of medications changes over time. In particular, for those patients who had 2 (or more) visits, we want to see whether the number of medications tends to be different from the first visit to second visit. Use the appropriate statistical test to investigate this.

Note that this problem requires a fair bit of data wrangling to get the data prepared. So consider these hints:

1. We can assume `encouter_id` is a variable that uniquely identifies each hospital admission.
2. We further assume that encounter IDs are only increasing, so if a given patient has two encounters, the first ID will always be less than the second ID.
3. The shape of this data set is typically called "long format"; the patients have repeat measurement recorded as new rows. You will want to get these data in to "wide format". There are good examples you can find by Google-ing.
4. Consider using the `dplyr` package. It's not necessary, but it might help.
5. The data reshaping is definitely tricky, don't feel bad if you struggle a bit. It's important to get practice with this, because for many of us, data cleaning and reshaping is a big part of the work we have to do prior to model fitting.