Hypothesis Testing & Central Limit Theory Exercise Solutions

January 10, 2019

1 Demonstration of the Central Limit Theorem

The exercises in this section help us examine the central limit theorem (CLT) in action. In particular, we see how sample sums taken from a Binomial tend to be normally distributed. We also get bit of an odd case with the Cauchy distribution, which will *not* behave the way many other distributions do with respect to the CLT.

1.1 Approximating the Binomial

This exercise wants us to write a function that simulates tossing `n` coins and taking the count of those that turned up "heads". Moreover, we want to be able to simulate this some arbitrary number of times; call this `n_sim`.

The function below simulates tossing `n` coins, and running the simulation `n_sim` times. The result is a vector of length `n_sim`, where each element in the vector is the number of times "heads" came up in the corresponding simulation. Note that we use the `rbinom()` function to simulate tossing `n` coins.

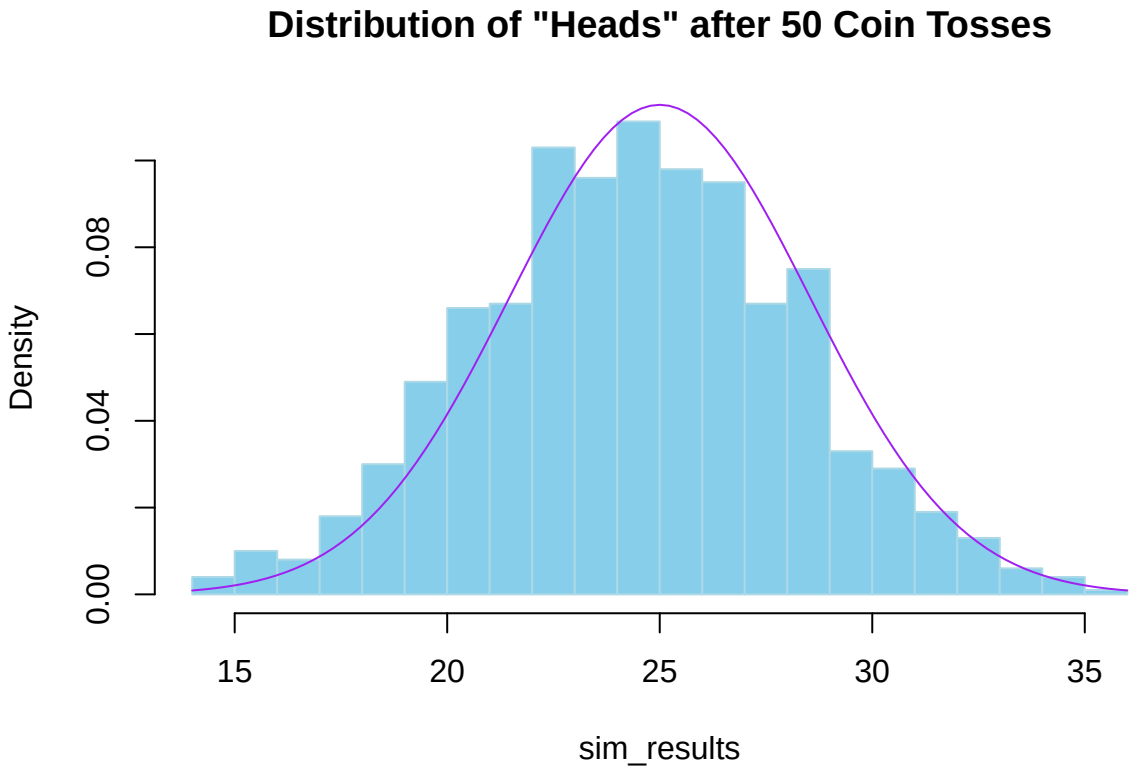
```
sim_binomial <- function(n, n_sim) {
    sum_arr <- rep(0, n_sim)            # pre-allocate array to store sums

    for (i in 1:n_sim) {                # iterate from 1 to n_sim
        tosses <- rbinom(n, 1, 0.5)     # take n draws from binomial distribution
        sum_arr[i] <- sum(tosses)       # sum of the draws
    }
    return(sum_arr)
}
```

We can then use this function to simulate tossing `n` coins `n_sim` times. Then we plot the resulting distribution using the `hist()` function. And for reference, we can overlay the normal distribution. The appropriate normal distribution has a mean (according to the definition for the binomial distribution) that is $n \times p$, where $p = 0.5$ in our case, and the standard deviation is $\sqrt{np(1-p)}$.

The plot below shows the distribution of sample sums for `n_sim = 1000` and `n = 50`. We can see from the plot above that the distribution of sample sums (i.e., occurrence of "heads") appears quite normally distributed. Furthermore, the larger the value of `n` and `n_sim` the more closely the approximation tends to become.

```
# run our function for n = 50 and n_sim = 1000
n <- 50
sim_results <- sim_binomial(n, 1000)
```

```r
# plot results
hist(sim_results,
     freq = FALSE,
     col = 'skyblue',
     border = 'lightblue',
     breaks = 30,
     main = 'Distribution of "Heads" after 50 Coin Tosses')
curve(dnorm(x, n*0.5, sqrt((n*0.5)*(1 - 0.5))), add = TRUE, col = "purple")
```



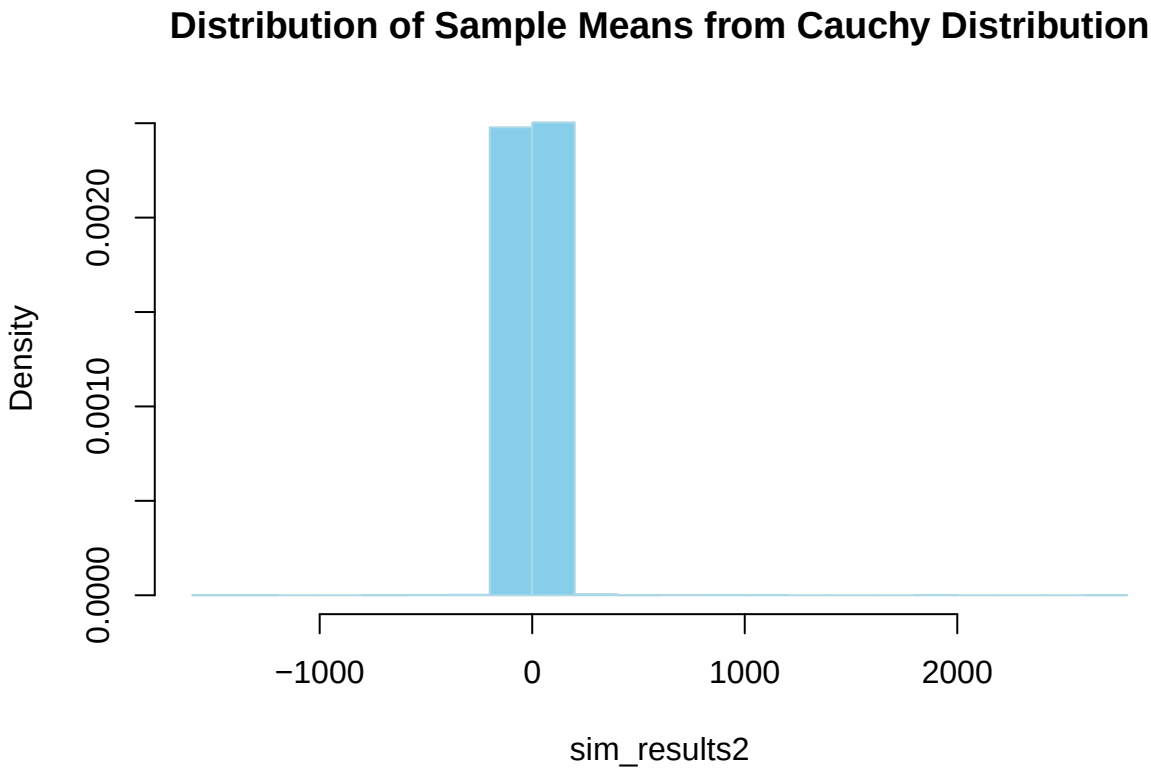
1.2 Sampling from a Cauchy

We can make some very minor changes to our function from above in order to take samples from a Cauchy distribution with location 0 and scale 1.

```r
sim_cauchy <- function(n, n_sim) {
    mean_arr <- rep(0, n_sim)           # pre-allocate array to store sums

    for (i in 1:n_sim) {                # iterate from 1 to n_sim
        sim_vect <- rcauchy(n, 0, 1)    # take n draws from Cauchy distribution
        mean_arr[i] <- mean(sim_vect)   # mean of the draws
    }
    return(mean_arr)
}
```

However, when we attempt to plot the results of the simulation, we note that the simulated sample means do *not* appear to be normally distributed. In fact, the reason for this is that the population mean and variance are undefined for the Cauchy distribution. And recall that the central limit theorem (CLT) works in cases where we have both a mean and variance that exist. The Cauchy is known as a "pathological" distribution. It does not satisfy the assumptions of the CLT, and rather than being normally distributed, the distribution of sample means from a Cauchy are themselves Cauchy distributed.

```
n <- 100
sim_results2 <- sim_cauchy(n, 10000)

# plot results
hist(sim_results2,
     freq = FALSE,
     col = 'skyblue',
     border = 'lightblue',
     breaks = 30,
     main = 'Distribution of Sample Means from Cauchy Distribution')
```



2 Diabetic Patient Admissions

The exercises here touch on a number of concepts, including using the binomial test to investigate the distribution of a sample relative to some hypothesized theoretical value. We also use the χ^2 test of indenpendence to examine whether or not two categorical variables are independent. Finally, we do a bit of data filtering in order to prepare our data for analysis.

2.1 Admissions by Gender

We can use a binomial test and observe that females are slightly over-represented in these data. In particular, females make up 53.7% of the sample. And since we have a fairly large sample (e.g., 100k), this is statistically significantly different from what we would expect if $p = 0.5$ in the population.

```
library(dplyr)
library(readr)

dia <- read_csv("diabetes_data_clean.csv")
```

```
is_female <- dia$gender == "Female"   # vector of bools for whether patient is female
n_females <- sum(is_female)           # number of success
n_patients <- length(dia$gender)      # number of trials
binom.test(n_females, n_patients)     # binomial test with p = 0.5 (default)
```

```
##
##  Exact binomial test
##
## data:  n_females and n_patients
## number of successes = 54708, number of trials = 101770, p-value <
## 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5345170 0.5406533
## sample estimates:
## probability of success
##              0.5375862
```

We can can use a χ^2 test of independence to determine whether or not the variables `gender` and `admission_type` are independent. As the results indicate, we do observe that the two variables, `gender` and `admission_type` are not independent. From checking residuals, it seems there were less female patients electively admitted than the null would have predicted, and there were more male patients than the null would have predicted. Additionally, it seems that the less male patients categorized as emergency admission than expected under the null hypothesis.

```
# First, we take a subset of the data to include only those
# admissions that are in the Emergency or Elective category.

dia_subset <- dia %>%
```

```
                filter(admission_type == "Emergency" |
                       admission_type == "Elective") %>%
                filter(gender == "Female" |
                       gender == "Male")

# Take a look at crosstabs
xtabs(~ gender + admission_type, dia_subset)
```

```
##         admission_type
## gender   Elective Emergency
##   Female     9840     29448
##   Male       9028     24540
```

```
# Run chisq test of independence
chi2_test <- chisq.test(dia_subset$gender, dia_subset$admission_type)

# Examine residuals
residuals(chi2_test)
```

```
##                  dia_subset$admission_type
## dia_subset$gender  Elective Emergency
##            Female -3.317879  1.961439
##            Male    3.589449 -2.121984
```

We can make a very small change to examine whether `admission_type` and `gender` differ for those patients referred by physicians. We simply take another subset from our existing subset. We still use a χ^2 test of independence to answer our research question. The results essentially mirror those observed above (where we didn't condition physician referral).

```
# First, we take a subset of the data to include only those
# admissions that are in the Emergency or Elective category.

dia_subset2 <- dia_subset %>%
                filter(admission_source == "Physician Referral")

# Take a look at crosstabs
xtabs(~ gender + admission_type, dia_subset2)
```

```
##         admission_type
## gender   Elective Emergency
##   Female     8399       932
##   Male       7787       745
```

```
# Run chisq test of independence
chi2_test <- chisq.test(dia_subset2$gender, dia_subset2$admission_type)

# Examine residuals
residuals(chi2_test)
```

```
##                   dia_subset2$admission_type
## dia_subset2$gender   Elective  Emergency
##            Female -0.6089595  1.8918714
```

```
##             Male    0.6368352 -1.9784737
```

3 Diabetic Medications

In these exercises we get to explore some continuous variables. We use few different t-tests to investigate mean differences. And we do some data wrangling to get everything in to shape for prior to analyzing the data

3.1 Medications by Admission

We can use an independent samples t-test to determine whether patients coming in electively or in the case of an emergency differ in their number of medications. The results indicate that we do indeed see a significant difference in the number of medications for patient who are admitted electively versus those entering in the case of an emergency. In particular, those patients who were admitted electively tended to have a larger number of medications.

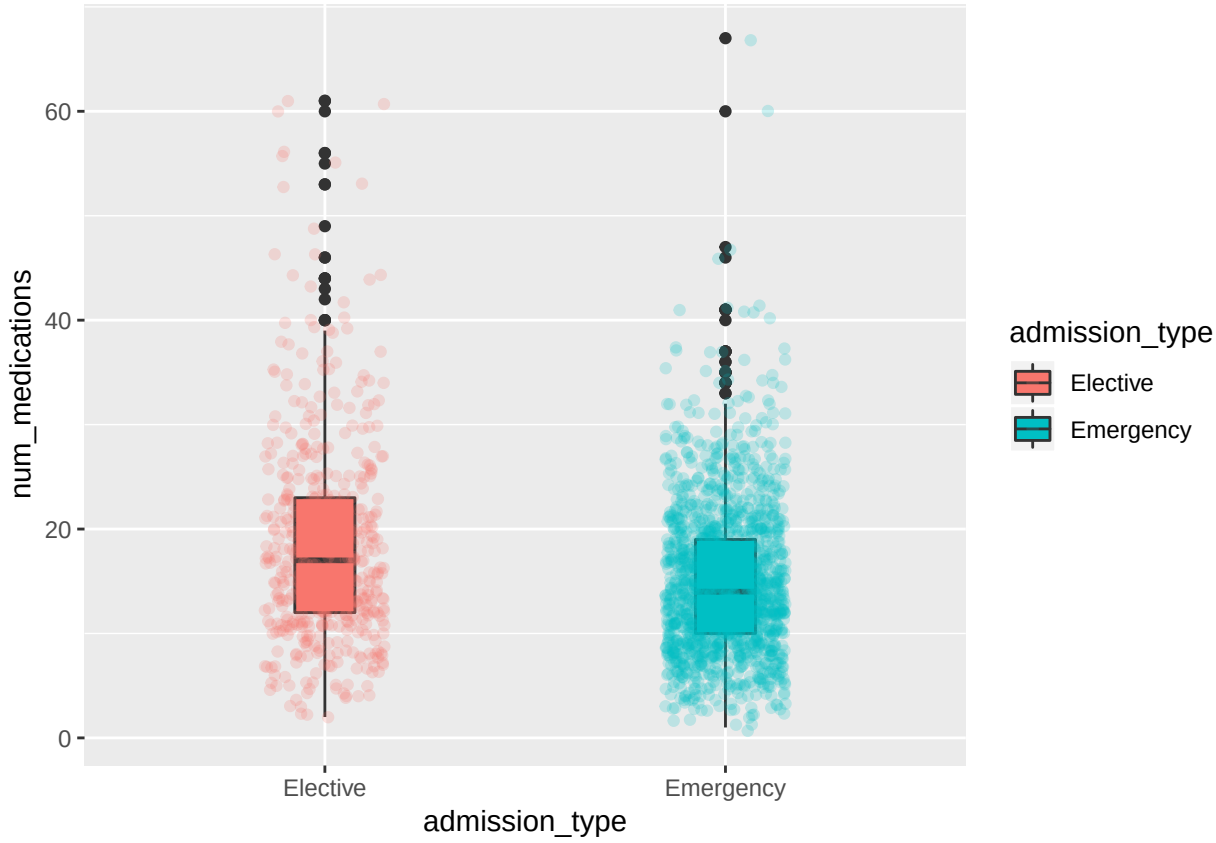
We should note also, however, that one of the assumption of the independent t-test is that the observations are independent. And in this case, we believe that to be false. A bit of investigation reveals that more than 16,000 patients had repeat visits, and thus they appear in the data set more than once. This violates one of our assumptions, so any inferences ought to be made with some caution.

```
library(ggplot2)

# First, take a subset of the data to include only those
# admissions that are in the Emergency or Elective category.

dia_subset3 <- dia %>%
                 filter(admission_type == "Emergency" |
                        admission_type == "Elective")

t.test(num_medications ~ admission_type, dia_subset3)
```

```
##
##  Welch Two Sample t-test
##
## data:  num_medications by admission_type
## t = 41.774, df = 27028, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.096853 3.401774
## sample estimates:
##  mean in group Elective mean in group Emergency
##                18.63453                15.38522
```

```
# random sample of 2000 for plot
rand_sample <- sample(nrow(dia_subset3), 2000)

ggplot(dia_subset3[rand_sample, ], aes(y = num_medications,
                                       x = admission_type,
```

```
                                    fill = admission_type)) +
    geom_boxplot(width = 0.15) +
    geom_jitter(width = 0.15, aes(colour = admission_type), alpha = 0.2)
```



3.2 Medication and Repeat Visits

We can use a paired-sample t-test to investigate wheth the number of medications remained fairly stable for patients for those making repeat visits. As the result below indicates, we do find a significant difference, in that patients tended to have a higher number of medicatis on their second visits.

The most tricky portion is the pre-processing needed to prepare the data. In particular, we first have to sort the data to ensure that we will be able to get patients first and second visits. We then exclude any patients without 2 or more visits. Then we group on the patient ID (i.e., `patient_nbr`), and use the `summarise` verb from `dplyr` to create two new columns `n_meds1` and `n_meds2`. These have the number of medications for each patient at their first and second visits. Note that, as the name suggests, we are going from a long-format data set, to a wide format data set.

```
# Sort by encounter_id
dia <- arrange(dia, encounter_id)

dia_wide <- dia %>%
              add_count(patient_nbr) %>%        # add column with num visits for each patient
              filter(n > 1) %>%                 # include only repeat patients
```

```
            group_by(patient_nbr) %>%
            summarise(
                n_meds1 = num_medications[1],
                n_meds2 = num_medications[2]
            )

# Examine descriptives
summary(dia_wide$n_meds1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   11.00   15.00   15.95   20.00   72.00
```

```
summary(dia_wide$n_meds2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   11.00   15.00   16.33   20.00   68.00
```

```
t.test(dia_wide$n_meds1, dia_wide$n_meds2, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dia_wide$n_meds1 and dia_wide$n_meds2
## t = -5.1459, df = 16772, p-value = 2.692e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5204836 -0.2333470
## sample estimates:
## mean of the differences
##              -0.3769153
```