Lab Expected CLT

Abhi Thanvi, Paul Holaway

July 11th, 2022

Contents

Lab Expected CLT

Welcome

Just like learning a new spoken language, you will not learn the language without practice. Labs are an important part of this course. Collaboration on labs is **extremely encouraged**. If you find yourself stuck for more than a few minutes, ask a neighbor or course staff for help. When you are giving help to your neighbor, explain the **idea and approach** to the problem without sharing the answer itself so they can figure it out on their own. This will be better for them and for you. For them because it will stick more and they will have a better understanding of the concept. For you because if you can explain it to other students, that means you understand it better too.

The Idea of this Lab

The idea behind this lab is the importance of Central Limit Theorem and Random Variable. Central Limit Theorem (aka CLT) is the basis of lot of the statistical inference procedures that you will see later in this course. In other words, CLT allows us to do a lot of the statistics we do today. Random Variable is a tool for statisticians to approximate the distribution of our sample and calculate certain attributed about our sample. What we learned today and apply in this lab might seem like an abstract or "random" topic (haha random), but trust me that these topics are important for becoming a good statisticians, and eventually a even better data scientist.

"Randomness in statistics, if done many times and correctly, can show patterns!" - Woke Abhi

Problem 1: Central Limit Theorem (CLT)

Question 1: Among your group, discuss what is Central Limit Theorem? Report an answer in your own words.

Answer: The Central Limit Theorem states that any distribution, when independently and randomly sampled enough times (usually at least 25-30 times), will approximate the normal distribution.

Question 2: If you have a gamma distribution, and you sample it 1000 times. What distribution will it most likely follow and why?

Answer: Assuming that you sample this distribution independently and randomly, it will most likely follow (approximate) the normal distribution, because due to the Central Limit Theorem, when you randomly and independently sample any distribution enough times (in this case, 1000 is definitely enough samples), the distribution approximates the normal distribution.

Problem 2: Random Variable (RV)

Question 1: Let's suppose you are playing a game with a friend called Winner Takes All. You win the game if you roll an odd number, but lose an even number. What would the expected number and standard deviation of wins be if you rolled 3 times?

Answer: $P(win) = P(odd) = \frac{3}{6} = 0.5$
$n = 3; p = 0.5$
–
$\mathbb{E}(wins) = \mu_{wins} = np$
$\mathbb{E}(wins) = 3 * 0.5 = 1.5$
–
$\sigma_{wins} = \sqrt{Var(wins)} = \sqrt{np(1-p)}$
$\sigma_{wins} = \sqrt{3 * 0.5(1-0.5)} = 0.5\sqrt{3} \approx 0.866$

Question 2: A major IT company makes some really cool gadgets. One out of every 50 gadgets is faulty, but the company doesn't know which ones are faulty until a buyer complains. Suppose the company makes a *$5* profit on the sale of a working gadget. The downside is that the company loses *$60* for every faulty gadget because the customers return the product and they have to repair it. Will the company make a profit for long term?

Hint: What does Expected Value mean? What does Standard Deviation mean?

Answer: Expected Value, in this context, refers to the expected (or average) amount of profit per gadget that the company makes. Standard Deviation refers to the spread of this expected value.

Let X represent the profit that the company makes for each gadget sold.

$\mathbb{E}(X) = \mu_x = \sum_{\forall x} xP(x) = (5) * \frac{49}{50} + (-60) * \frac{1}{50} = 4.9 - 1.2 = 3.7$
–
$\sigma_x = \sqrt{Var(X)} = \sqrt{\mathbb{E}(X^2) - (\mathbb{E}(X)^2)} = \sqrt{\sum_{\forall x} x^2 P(x) - \mu_x^2} = \sqrt{(5^2 * \frac{49}{50} + (-60)^2 * \frac{1}{50}) - (3.7^2)} = \sqrt{96.5 - 13.69} = 9.1$

Because the expected value is 3.7, we are assured that the company, on average, will make $3.70 per gadget, and will therefore make a profit long-term. Although the standard deviation is 9.1 (which seems huge), over a long period of time, the Law of Large Numbers assures us that the average profit will get closer to the expected value, so on average, over a long period of time, the company makes a profit of 3.70.

Question 3: The Elkhart and Western is a Class III short line in northern Indiana. However, they are running more trains than they used to. John, who lives next to the tracks and works from home, is trying to figure out how many trains he should expect to disturb him. The number of trains that passes by his house and the probabilities are listed in the table below. Calculate the expected value and standard deviation of the trains that pass by his house.

Trains	Probability
0	0.35
1	0.10
2	0.30
3	0.05
4	0.20

Answer:
$\mathbb{E}(Trains) = \mu_{Trains} = \sum_{\forall x} xP(x)$
$\mathbb{E}(X) = 0(0.35) + 1(0.10) + 2(0.30) + 3(0.05) + 4(0.20) = 1.65$
–
$\sigma_x = \sqrt{Var(X)} = \sqrt{\mathbb{E}(X^2) - (\mathbb{E}(X)^2)} = \sqrt{\sum_{\forall x} x^2 P(x) - \mu_x^2}$

$\mathbb{E}(X^2) = 0^2(0.35) + 1^2(0.10) + 2^2(0.30) + 3^2(0.05) + 4^2(0.20) = 4.95$
$\mu_x^2 = 1.65^2 = 2.7225$
$\sigma_x = \sqrt{\sum_{\forall x} x^2 P(x) - \mu_x^2} = \sqrt{4.95 - 2.7225} = \sqrt{2.2275} \approx 1.492$

Question 4: Import the Data.csv and calculate the E(X) and SD(X) of Siblings.

Answer

```
#Import the Data set
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
Data <- read.csv("~/Desktop/DPI 2022/Data.csv", stringsAsFactors=TRUE)
#E(X)
print("Expected Value of Number of Siblings:")
```

```
## [1] "Expected Value of Number of Siblings:"
```

```
mean(Data$Siblings.)
```

```
## [1] 1
```

```
#SD(X)
print("Standard Deviation of Number of Siblings:")
```

```
## [1] "Standard Deviation of Number of Siblings:"
```

```
sd(Data$Siblings.)
```

```
## [1] 0.6900656
```

Project Questions

Feel free to work on your project if there is any time left after the labs. Paul and I are here to answer any questions during the second half of the lab times to answer mainly project related questions, but general questions are more than welcome too. Feel free to discuss among your group about any project ideas or help each other out. Remember collaboration is promoted, plagiarism is not! :)

Submission

Once you have finished your lab...

1. Go to the top left and click `File` and `Save`.
2. Click on the `Knit` button to convert this file to a PDF.
3. Submit **BOTH** the `.Rmd` file and `.pdf` file to Blackboard by 11:59 PM tonight.