Dustin McDonnell

**Indicators of Depression**

The dataset I chose to preform my calculations on is a report by the U.S census Bureau originally produced to obtain data on the social and economic impacts of covid-19 on American households. The dataset I am using specifically studies indicators of anxiety or depression based on symptoms measured over weekly timespans. Although started when covid was in full swing the study continued until March 2024 and the scope of demographics of the study grew as time went on. When I first was browsing possible data sets and sorted this study by states I saw the population of Florida skewed more depressed than most states. I was stationed in Florida for a few years, so this fact makes complete sense to me. In addition, I just found the study humorous, so this is what I settled on.

One of the first things I did was set up my csv as a table in excel for ease of use, for the most part I won’t have to worry about simple math this way and can focus on the probability portion. A picture of this is shown below, I have a few filters in place so the entirety of the table can be seen the actual dataset is much larger.

A screenshot of a computer

Description automatically generated

For this dataset I am making a few assumptions that may not be correct, but the survey says weighs responses to account for the entire population by age, gender, race, and ethnicity. So, for the sake of this report, I will be assuming the total population of the United States is 333.3 million people. When I perform operations on demographics I will be making similar assumptions, and I will mention what numbers I am using prior to the operations.

For some basic set theory as a warmup, we will denote D as a depressed person going to a grocery store and H is a happy person going to a grocery store.

is the event that either the happy or the depressed person make it to the grocery store.

is the event that both the happy and depressed person make it to the grocery store.

means neither of them make it to the store.

I am aware that this did not actually need any math, but for the sake of completeness I included this earlier section of the book.

There are five subgroups of ethnicity tracked in this study. Non-Hispanic Asian, single race, Non-Hispanic White, single race, Non-Hispanic Black, single race, Hispanic or Latino, Non-Hispanic, other races and multiple races. I do think the fifth category is a little broad but that’s what we’re working with. There are more subcategories in this study but for the moment this is what we will be working with. The total number of the population is not listed in the data set, but some research online lists the populations as follows:

Shorthand:

Single Race= S

Mixed Race = M

Hispanic = H

Non-Hispanic = N

N Asian S = 18.9 million / 333.3 million = 5.6 % of the population

N White S = 197,639,521 / 333.3 million = 59.3 % of the population

N Black S = 40.1 million / 333.3 million = 12.1 % of the population

Hispanic S = 63.7 million / 333.3 million = 19.1% of the population

N Multiracial = 12,192,866 / 333.3 million = 3.67% of the population

Which all sums up to around 99.77 percent of the population. These are mixed numbers from different studies but that’s pretty close to 100 so that’s what we’re working with.

The timespan for the data I am using is around four years so any numbers I use going forward are only accurate for that time period.

***Self-Reported Indicators of Depression/Anxiety on Average over the entire study period:***

I am getting these averages from the excel table like so:







Total:

A screenshot of a table

Description automatically generated

These are not all the values because that would be a prohibitive amount of space, I am just showing the categories. The values are the self-reported rates of depression every week averaged over the time period.

N Asian S = 18.9 million \* 27.1123% = 5124224 people displaying symptoms

N White S = 197,639,521 \*32.8% = 64825762 people displaying symptoms

N Black S = 40.1 million \* 35.8% = 14355800 people displaying symptoms

Hispanic S = 63.7 million \* 38.6& = 24588200 people displaying symptoms

N Multiracial = 12,192,866 \* 43.88% = 5350229 people displaying symptoms

A lot of people in America self-report depression/anxiety symptoms, it’s especially painful as a mixed raced person myself to see the rate is so high.

This isn’t relevant to the data but lets say there are 60 people and 3 of them are depressed. That would be and of those 60 people 10 of them are multiracial, the probability ignoring the numbers above would be / = .0035. This has nothing to do with the percentages above I just thought it was a grim hypothetical.

On my excel chart it has the total rate of negative symptoms among men and women is as follows:

The rate amongst men is P(D|M) = 30%

The rate amongst women is P(D|F) = 37%

If we assume P(Being a Man) = P(M) and P(Being a woman) = P(F) and both are around .5 for simplicities sake we can calculate using conditional probability.

P(D|M) \* P(M) = = .3 \* .5 = .15

And if we preform the same calculation for women using their respective numbers it ends up being .185.

This is consistent with online resources.

A screenshot of a cell phone

Description automatically generated

With the numbers trending slightly higher in this study. However, this study also does not track being diagnosed with depression just negative self-reported symptoms so it makes sense that the study would trend higher.

We can perform a fun hypothetical using Baye’s theorem. Let’s say a STEM classroom contains unfortunately 90% men and 10% women due to systemic inequality. Knowing that P(D|M) = 30% and P(D|F) = 37% we can determine the conditional probability of choosing a woman with depressive symptoms from the room.

Which comes out to 12%, which is surprising to me? Maybe I did it wrong. I just thought it would be lower.

Here’s another hypothetical, the rate of report on anxiety/depressive symptoms for those with less than a high school diploma is frighteningly high at 41.5%





We could calculate that in a GED office filled with 20 people the odds that at least one person is displaying negative symptoms.

P(at least one with negative symptoms) = 1 - = .99999997

It can be seen that education is important and should be accessible to everyone.

Americans over the age of 80 have a lower rate of negative symptoms than any other demographic, I imagine time acts like a filter of sorts. By the time you hit that age you probably have it figured out.





So, let’s say a grandchild is told by her grandparent that her friend in the old folk’s home is in a low mood and needs cheering up, but the child doesn’t actually know what they look like when the grandparent shooed them away to cheer the friend up. Thinking quickly the child enters the common area and to their shock all the old people look sad. The ravages of time have made this whole operation a lot harder. Luckily old people are forgiving of the mistakes made by children, and we can figure out the odds the child correctly guesses which old person is the friend of their grandparent on their third guess like so:

***Geometric***

Q = .1815

P = 1- .1815 = .8185

P(3) = 12%

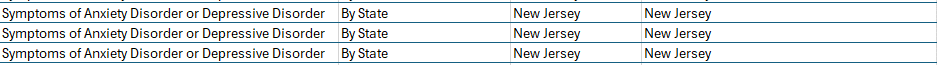
This is unfortunate for the child; they have a very low chance of actually finding the friend before some cantankerous geriatric gets upset with them.

We could use the same scenario as above but instead of the third person being the first negative symptom bearing pensioner what if that person was the second of them who had negative symptoms?

***Hypergeometric***

P(3) = = .199%

These are slightly better odds, but the unfortunate reality is that just means the child found more sad people and it might not have been the person they were looking for in the first place.

The average rate over a time period of three weeks in New Jersey for self-reported symptoms is 34.9%.

A screenshot of a calculator

Description automatically generated

So, let’s say that instead of 3 people out of ten self-report on average per week but over that time period 9 people self-reported. Does this seem plausible? We can find out with Poisson distribution.

***Poisson distribution***

P(9) = = .0027

Not only is the scenario described incredibly unlikely it’s borderline impossible.

The rate of depressive symptoms sorted by gender identity amongst the transgender community is distressingly high sitting at 71.1213%.  




Out of 100 transgender individuals can we figure out the expected number of people displaying depression signals. Along with this if we arbitrarily pick 20 as the amount of acceptable people per 100 to be struggling with mental health for a healthy society would we remain under that metric?

Y has a binomial distribution of n = 100

P = .71

Q = .29

The expected number of people to display symptoms would be 71 from 71% \* 100, that was easy because we used such clean numbers.

We can calculate the expected value of n with:

And we can calculate the variance with:

And the standard deviation with:

= 4.54

To get the upper and lower bound:

And finally, by Tchebysheff’s theorem:

= .75

There is a 75 percent chance that in a group of 100 transgendered individuals that the amount of individuals displaying depressive symptoms in the bounds of (. This is tragic.

Just to be clear, this was not intended to disparage anyone in the LGBT community I’m just preforming calculations on my spreadsheet and frankly I hope I got this part wrong because the number is way too high.

From this point forward I imagine that math is going to be completely wrong, I tried.

***Continuous Variables***

Going all the way back to the start of this report where I got the averages of depression amongst the different ethnicities. I will be using the mixed rate statistic of self-reporting which sits at an average of 44%.

Let’s say 3 random mixed-race individuals are selected, one of them may be displaying depression symptoms. A therapist checks them one at a time.

We can identify the probability of the first depressed person being identified on Y as such:

= .290

= .44 \* .66 \* .66 = .195

Following this we can map the distribution function as:

From this we can get the probability of things such as P(Y2) = P(Depression symptoms identified at 2 or less individuals).44 + .290 = 73%

From the density function we can get the expected by:

Which comes out for me as:

Which, I do not think can actually be an expected value. I don’t think I did the integral part wrong. I’m actually fairly certain that the previous section with the distribution function is the error, and I’m probably going to fail the final, so I don’t know for sure where the issue is.

***Uniform Distribution***

I parsed my data for quite some time, but unless I am severely misunderstanding uniform distribution the nature of depression symptoms makes them not uniform. There does not seem to be any period of time in any of the demographics over any time period where the rate of symptoms is stable enough to be uniform. Which is a shame because I actually know how to do this one.

Let’s just say hypothetically there exists a small town in the Midwest where the report rate exists uniformly between 30 and 40 percent per week. We could find the probability that the next week’s report week is let’s say under 34 by doing the following:

= .4

Making the probability 40 percent.

The mean and variance are also easy to get from this,

Mean:

= 35

And

At this point in the report, I have reached a conundrum. I no longer have the mathematical ability to do the rest of this report justice assuming I even did most of the parts before this correctly. Do I press on knowing it’ll be incorrect? Does that devalue the rest of this report? Will I forget to cut this part out?

***Gamma Probability Distribution***

From my slim understanding of Gamma probability distribution, it’s used for skewed probability density functions where most of the area is near the origin. The example used in the book is the length of time between aircraft malfunctions. The rate at which depression symptoms happen rise significantly more around June for some reason so that could be something similar in terms of skewed area?

I’m not one hundred percent clear on where the domains of -infinity to infinity come from in terms of integration for Gamma distributions.

So, to this end I’m choosing this week.



It’s June in Florida with an average report rate of 43.8 and let’s just arbitrarily pick 100 people so the mean is 44.

This would mean = 44.

We can determine the probability that P(more than 45 people self-report) for the next week.

If I understand this correctly we would perform this operation like so:

Which integrates to

= 36% chance that more than 45 people self-report.

I am aware the setup for this problem is probably completely wrong and they probably will be going forward, but I hope the execution of the problem is correct.