Menstrual Cycles & Their Predictive Power:

Predicting Period Cycles & Illness

Jessica Chudnovsky

The interactive website I created based on my models can be found here: link

Introduction

Nearly all females have it: a menstrual cycle. Despite it being an entirely natural process, it is rarely talked about. When we imagine heart illness, we think about cholestoral levels. When we imagine viruses, we look at symptoms like fever, chills, and congestion. Given the taboo nature of women's health, rarely do individuals feel comfortable sharing details of their cycle, its length, and understanding the data and what it means. Outside of annual health checkups, periods are an entirely taboo topic. (Fun fact: when I told a friend of mine that I wanted to do my CS 109 project on Menstruation, they laughed, and were concerned that no one would be able to take this topic seriously given how unusual it is to speak openly about it.) There's not a lot of publicly available data, and there is a lack of adequate data in order to determine how it correlates with other health issues. **That's why this project builds a user interface for women to enter very simple data about their cycles as well as other symptoms and illnesses they face, in order to be able to make better predictions over time about their relationship. Link: here.**

You may ask: so what? You understand that it's a natural female process, but so is eating and breathing. It doesn't mean that it has to be a focal point of discussion. However, perhaps it should be. Periods and statistics about them are an incredibly important indicator for many illnesses and conditions that females face, such as PCOS (Polycystic ovary syndrome), which is a condition that 6 to 12% of women of reproductive age face,[1] as well as Ovarian Tumors, and other conditions that impact reproductive and overall health and wellbeing of women. With heightened anxiety levels or sleep deprivation, periods become more irregular.

Periods are a symptom of very serious illness such as PCOS and Ovarian Tumor, which are illnesses that are often misdiagnosed, leading to not being treated for years. This project aims to leverage period cycle data in order to make assessments about irregularity and regularity, the former judgement of which often correlates with larger health issues that may need to be diagnosed. This is determined through a bootstrapping method of means of regular period data compared with real data, which calculates a p value in order to determine regularity or lack thereof. Then, it aims to predict the likeliest next date of period depending on the regularity or irregularity of the period, using a combination of a Normal And Weibull distribution.

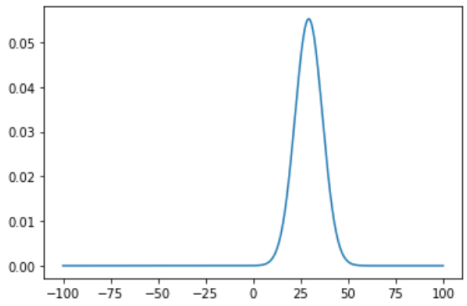
My own inspiration for this project was a good friend of mine (who's data is leveraged in the project) first suffered from irregular periods, and it took so much time to properly diagnose her with PCOS. However, there was an even deeper underlying issue: an Ovarian Tumor. Her symptom of incredibly irregular periods was overlooked for so long, and the same is true for

conditions like Uterine Fibroids, which members of my family suffer from as well. We don't ask enough questions about this issue, which is directly the problem that this project addresses.

How to make deductions from a period

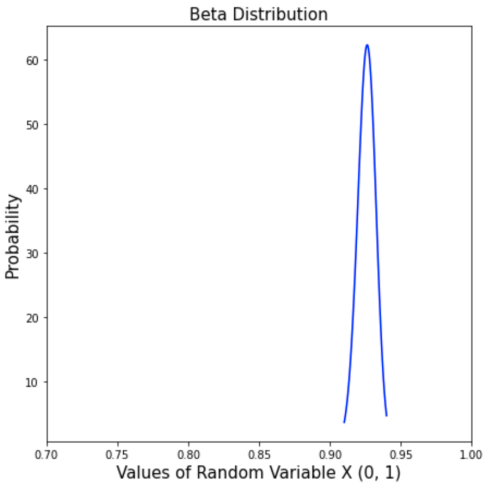
One of the more standard measures of a period is cycle length: the number of days from the first day of one period to the start of the next period. A cycle is considered abnormal if the cycle is less than 21 days or greater than 35.[2] We cannot however simpy conclude that a woman's cycle is irregular based on one individual cycle being irregular, hence the methods employed by this paper.

Guassian of how healthy periods are distributed



Understanding our data through Beta

I leveraged a dataset which we classify as majority "regular" cycles. From it, we can represent "regular" periods with the Beta distribution's parameter a and "irregular" periods with the Beta distribution's parameter b based on irregularity being determined by a cycle being outside the range 21 to 35, and a distribution of Beta(1545, 124) was obtained, which is an overrepresentation of the "regular" periods given that we calculate the maximum a posteriori (MAP) estimate as 0.92 from the calculation (a-1)/(a+b-2) = (1545-1)/(124+1545-2). However, we would estimate this parameter to be closer to 0.75 given that 25% of women have irregular periods, thus this dataset is classified as majority "regular," and we can use it as a baseline for regular periods.



The Beta distribution based on our dataset.

Assessing regularity and irregularity of inputted periods

We take as an input 12 data points of the user (though we accept less and encourage more)! The data points are the last 12 cycle lengths of the user, defined as the amount of time from the start of one period to the start of the next period.

```
def determine_regularity_mean(data):
  mean_cycle = np.mean(data)
  OBSERVED_DIFF = abs(mean_cycle - MEAN)
  n_greater_than_observed = 0
  grand_list = data + BASELINE_DATA

  for i in range(ITERATIONS):
    resample_1 = resample(grand_list, len(data))
    resample_2 = resample(grand_list, len(BASELINE_DATA))
    mean_resample_1 = np.mean(resample_1)
    mean_resample_2 = np.mean(resample_2)
    diff = abs(mean_resample_2-mean_resample_1)
    if diff >= OBSERVED_DIFF:
      n_greater_than_observed += 1
  p = n_greater_than_observed/ITERATIONS
  print("The p value for the mean is", p)
  return p
```

By leveraging the Bootstrap algorithm, we determine the observed difference between our regular data set and the inputted values, then calculate the probability of exceeding this observed difference by resampling from the combined data. Here is an example of what such data points look like:

32 28 46 16 23 15 17 28 25 26 28 27

After many helper functions allowing us to properly process and sample, simulate, and model this data, we're able to make the final calculation determining whether or not the period is regular based on our bootstrapped p values, and have 0.05 as our threshold for determining significance.

Bootstrapping for P Values, and Determining the Regularity of Period Cycle Based on Likelihood of Exceeding Observed Difference

```
def irregular_or_regular(data):
  p = determine_regularity_mean(data)
  if (p < 0.05):
      return "irregular"
  else:
      return "regular"
```

Testing our Data

Next, we put our assessments to the test!

First, we test it on simulated data. I generate random variables according to a Gaussian distribution that is formulated upon the mean and standard deviation of the original data set. Then, our irregular_or_irregular function determines whether the data is regular or irregular, and we add these outputs to our Beta in order to update our posterior belief over time.

Next, we are assessing the accuracy of this method by testing it on real data. To put my work to the test with its first user, I tested it on my data – I am confirmed not to be diagnosed with PCOS, Ovarian Tumor, or other health issues that correlate with irregular periods. My period data was correctly attributed to be regular. I also tested data against my closest friend who is diagnosed with an Ovarian Tumor and PCOS, and it accurately determined her data to be irregular. The scope of this project can and must be expanded to far more data points in order to eventually build a more reliably tested predictive model, which can be used to predict reproductive health issues. The model already effectively predicts with 90% accuracy based on labeled data collected from fellow Stanford students. (Based on a very small sample of 10 students. As it will be discussed below, proper data collection is an issue in this area.)

```
Here is an example 12 cycle lengths: [25.362125872409266, 22.068132456694443, 48.88212766604675, 16.910029838619693, 30.246840698620016, 30.706556686273586,
This pertains to the simulated_data:
The p value for the mean is 0.7091
The period is predicted to be regular

Here is my period data which we know is regular: [26, 21, 31, 33, 27, 29, 31, 29, 30, 23, 31, 23, 33, 27, 20, 25, 28, 33, 27]
This analysis pertains to my baseline data:
The p value for the mean is 0.0829
The period is predicted to be regular

Here is the period data we already know is irregular: [14, 37, 25, 42, 35, 55, 20, 23, 55, 50, 30, 35]
This analysis pertains to the irregular data:
The p value for the mean is 0.0
The period is predicted to be irregular
```

The Next Step: Determine the Likeliest Next Period Cycle Length

Now that we've been able to effectively label the inputted data as regular or irregular, our goal is to **Determine the Likeliest Next Period Cycle Length,** by using a combination of a Weibull and Normal Distribution, weighing each differently depending on whether the cycle is Normal or Irregular. We weigh the following two distributions, and from it, determine the likeliest next period cycle length. The Normal Distribution is simply some mean added to the Standard Normal, and the Weibull Distribution is a logarithmic function determined by certain constants dependent on the distribution of the data. This is inspired from theoretical research in the paper "Modeling menstrual cycle length using a mixture distribution."

Standard Cycles: Leverages a Normal Distribution

$$Y_{ij}=\mu+\mathrm{e}_{ij}^{(1)}$$

Non Standard Cycles: Leverages a Weibull Distribution

$$\log(Y_{ij}-s)=-\log\rho+\frac{1}{\kappa}\,\mathrm{e}_{ij}^{(2)}$$

Combined Distribution: Assignes weights to p and determines likeliest next day

$$f(y)=pf_1(y)+(1-p)f_2(y)$$

Results:

The first output corresponds to the next period cycle length predicted for me, and the second for my friend with extreme cycle irregularities. It accurately predicted the length within one day for both of us.

```
The predicted length of the next cycle is 30.77 days.
The predicted length of the next cycle is 38.90 days.
```

Here is the code

```
# Load data into a pandas dataframe
data1, data2 = personal_csv_reader()

# Define the parameters of the model
p = 0.5 # weight for Normal distribution
s = max(data1) # shift parameter for Weibull distribution
rho = 1.5 # shape parameter for Weibull distribution
kappa = 0.5 # scale parameter for Weibull distribution
mu = np.mean(data1) # mean of Normal distribution
sigma = np.std(data1) # standard deviation of Normal distribution

# Define a function to calculate the probability density function of the cycle length
def cycle_pdf(y):
    f1 = norm.pdf(y, mu, sigma)
    f2 = weibull_min.pdf(y - s, rho, scale=kappa)
    return p * f1 + (1 - p) * f2

# Define a function to predict the next cycle length based on the previous cycle lengths
def predict_next_cycle(data):
    x = np.array(data)
    output = irregular_or_regular(data)
    if output == "regular":
        return norm.rvs(loc=mu, scale=sigma) # sample from Normal distribution
    else:
        y = np.log(x[-1] - s)
        z = np.random.exponential(scale=kappa)
        return np.exp(y - np.log(rho) + z) + s # transform back to original scale

# Example usage
#cycle_lengths = [28, 31, 29, 30, 35] # previous cycle lengths
next_cycle1 = predict_next_cycle(data1)
next_cycle2 = predict_next_cycle(data2)
print(f"The predicted length of the next cycle is {next_cycle1:.2f} days.")
print(f"The predicted length of the next cycle is {next_cycle2:.2f} days.")
```

Why does this matter & applications

From numerous personal anecdotal evidence, I see time and time again that women who have irregular periods do not even track their periods in the same ways that women with more regular periods – they don't see the point, because it can't be of adequate use. Integrating these techniques for labeling different types of irregularities (and a potential extension is developing more models for different types of regularities) allows women with atypical period cycles to gain control over their cycle and develop more convenience in their lives. Something that occurs so often, and is such an integral part of a woman's health and reproductive system should not serve as such a burden. Further, we should be keen on determining the signals that period data guages for us, because these simple statistics give us the power to diagnose health issues such as PCOS, can be indicators for even more extreme health issues, and brings attention to treatments and getting the help that is needed. Issues like these impact reproductive health, and decrease the likelihood of fertility. If we can look at data points as simple as cycle lengths in order to properly diagnose these issues, it's vital that we don't overlook the power that we hold with this data. Menstrual cycles, a taboo topic we rarely

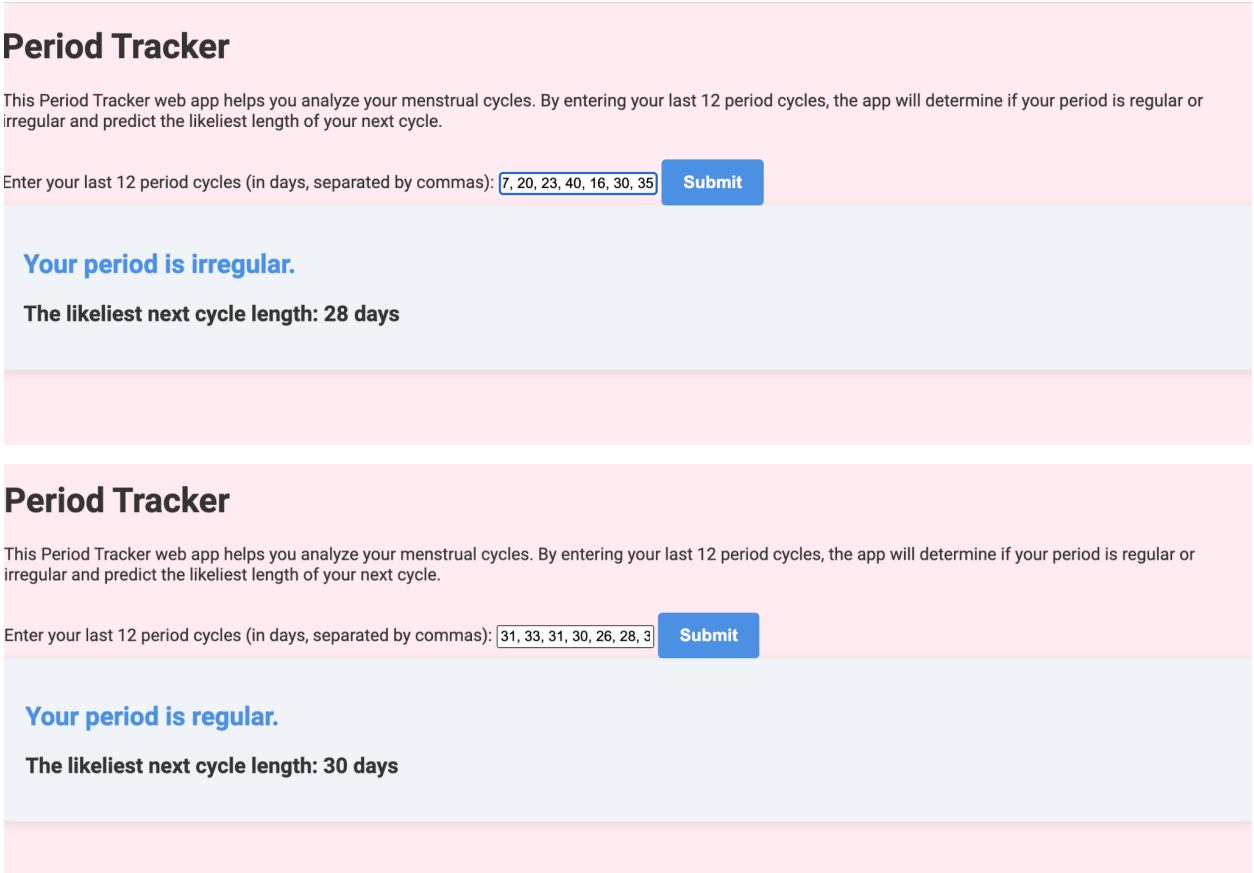
discuss, can often be an indicator for issues that prevent fertility, and menstrual cycles change the fabric of how women live their lives and their health.

However, one of the biggest challenges I experienced with this project was the lack of data. It was a difficult choice to make to work on this project, because I knew that I'd have to get creative with how I'd obtain and test my data – I was lucky to even get one dataset. Period data is so important for gauging metrics about womens' health, and having more publicly available anonymized datasets would create significant progress in female health. This project helped me realize that in a very vivid way. I'm also glad that I was able to use the opportunity to simulate data in a way that allowed me to test the tools I built.

Periods are an incredibly taboo topic, and it's often hard for women to confide amongst each other about these issues, let alone to a degree that would allow them to properly understand and escalate the issues that they face. The general lack of knowledge and awareness of these issues could potentially serve as a large risk for misdiagnosing reproduction related illnesses, as was the case with my closest friend who barely got diagnosed with her Ovarian Tumor in a timely enough manner.

I'm excited to be bringing attention to women's health issues through this project – often, discussion of these issues is limited due to its taboo nature, but this precise fact is a harm to society and women's health. It was also fun to be able to leverage my own data as well as close friends' of mine to see the impacts of my work firsthand. I'm hopeful that this project raises alarms for all of the insight we can gain from period data, and advocates for a need to obtain more publicly available data to help women effectively maintain their health with respect to their reproductive system and more.

Interactive Website of these Period Insights: Link



Conclusion

Period data can be guaged in order to draw correlatory data that allows one to gauge that a patient is at higher risk of reproductive issues that stem from PCOS and even Ovarian Tumor. Regular and irregular cycles can be classified and these classifications can be leveraged in order to develop predictions for the likeliest next period date.

Sources

[1]PCOS (Polycystic Ovary Syndrome) and Diabetes

[2]Irregular Periods

Modeling menstrual cycle length using a mixture distribution