EPIDEMIOLOGY REPORT: PY2.7 & PY3

The recent discovery of two strains, Py2.7 and Py3, amidst an influenza outbreak in Minervopolis has incurred uncertainty and unsettlement among our citizens in the past few days. In response, after working tirelessly to characterize these new strains and assess their acuity, Minerva lab has come to the following conclusions, addressed in the two sections below.

1. Due concern about new strains and their differences:

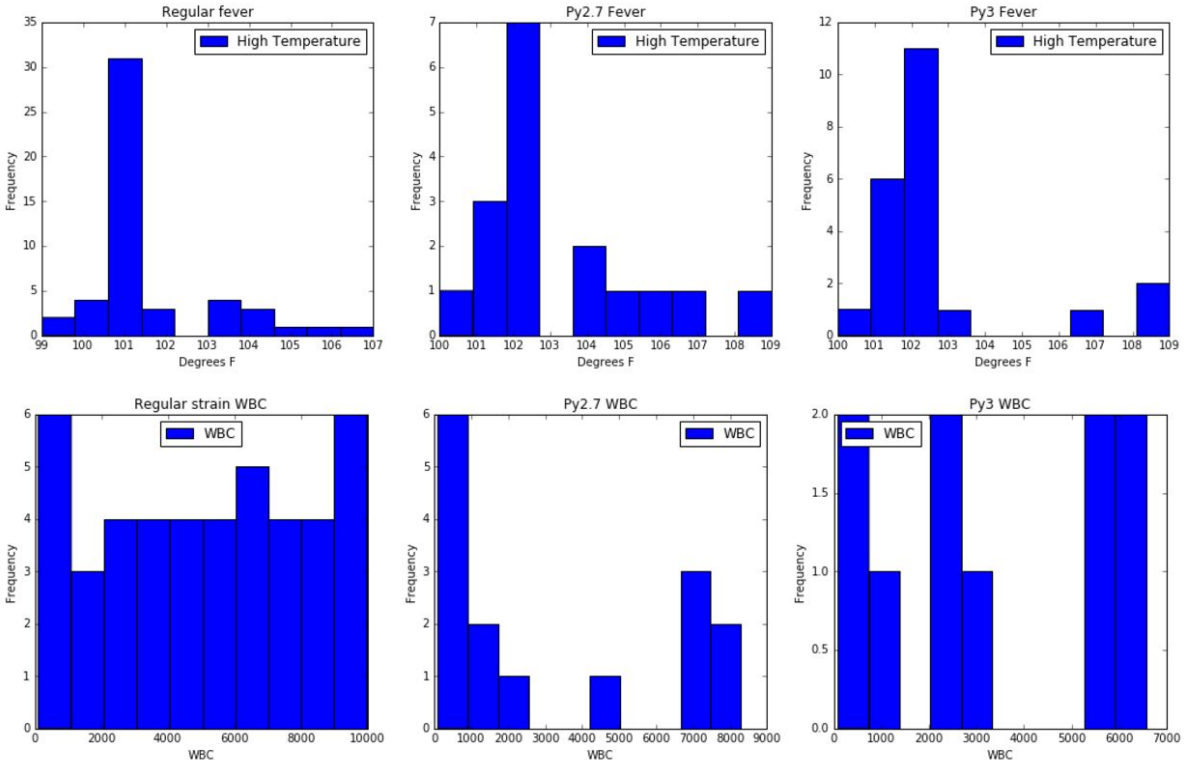
The initial data, although scarce (see appendix E), suggests that Py2.7 and Py3 are more acute than the regular influenza in terms of fever caused (effect size > 0.6 - See appendix B). Additionally, even though observation suggests that Py2.7 is more dangerous, with more cases of 104 degree F or above fever (35.29% for Py2.7 versus 13.63% for Py3), calculations prove that Py2.7 and Py3 are actually not statistically different (see Appendix A).

While white blood cell count of Py3 patients are more scattered, white blood cell count of Py2.7 patients suggest a high probability of secondary infection (see Appendix C). Both strains, however, seem to result in much lower white blood cell count than regular influenza.

See more suggestions of implications in appendix E.

2. Relation between high fever and new strains:

Calculations reveal that, in the current circumstance, 65.54% of patients with a fever above 102 degrees F would likely be suffering from either Py2.7 or Py3. This, however, might not the optimal way to self-diagnose as fever temperature is just one of the new strains' characteristics and it is easily subject to confusion of inverse fallacy. (see appendix D).



Pic 1: Histograms for the distributions of fever and WBC for regular strains, Py2.7 and Py3

APPENDIX A: HYPOTHESIS TESTS - FEVER

To address the current concern about the severeness and fundamental difference between the two strains, three hypothesis tests are conducted:

Py2.7 vs Regular strains	Py3 vs Regular Strains	Py2.7 vs Py3

These three hypothesis tests share the same **null hypothesis**, which is "There is no difference between (the mean of) the concerned subjects" (or $\mu_A = \mu_B$).

The corresponding **alternative hypotheses** for such null hypothesis would be: "There is a statistical difference between (the mean of) the concerned subjects" (or $\mu_A \neq \mu_B$).

1. Assumptions and concerns:

Nevertheless, first and foremost, we must **assume** that our samples (or subjects) are:

- Independent data in each sample and between samples (which are true as given by the problem, no one suffers from both Py2.7 and Py3 at the same time)
- Their distribution follow a normal distribution (which is not true, as they follow a lognormal distribution because of a natural limit - it would not be considered a fever if body temperature does not exceed 100.4 degrees F (emedicinehealth.com, 2017), especially when we only measure the patient's highest fevers - this would be addressed in Appendix E, but in the meantime, consider the distributions normal)
- Their standard deviation is approximately the same (which is acceptable in this case)

The Python Code (see Appendix F) returns the following **descriptive statistics** about the three samples of our concern:

Stats\ Samples	Regular	Py2.7	Py3
Sample Size	50	17	22
Mean	101.59	103.06	102.80
Standard Deviations	1.54	2.44	2.32
Min/ Max	99/107	100/109	100/109

To **test our hypothesis**, we would need to calculate the difference of two means, calculate their t-scores, and converse t-scores into p-value. Additionally, a confidence interval would ascertain the difference between two-means and quantify them with a certain rate of confidence.

The result (p-value) would then be compared with our **significance level** α to determine statistical significance. The standard α would be 0.05 (indicating a 5% chance of making Type I error). In the case of multiple hypothesis test like this one, a **Bonferroni correction**, in which the α of each test would equal 0.05 divided by the number of tests. However, I would consider this a too extreme a measure to apply in this particular circumstance (sample size is too small - see appendix E), in which extra care is more favorable than reckless overconfidence. It would be better to conclude that we have evidence/ strong evidence regarding the difference between the

new strains and regular strains than to outright reject the difference in case p-value is just slightly larger than the Bonferroni-corrected α*. Thus, I propose that we maintain a 0.05 α in this case.

T-test or Z-test: as the sample sizes for both Py2.7 and Py3 is under 30, and we do not know the population mean, but rather just a benchmark mean of 50 regular strain cases, it is sensible to use t-test rather than z-test.

One-sided or two-sided: We choose to conduct two-sided tests because the uncertainty of the situation requires more insightful discovery into the nature of the new strains.

***2*. *Calculations and results*:**

Take the hypothesis test between Py2.7 and regular strains for example. We have the difference of means $M_{2.7} - M_{reg} = 1.47$.

Standard error SE = $\sqrt{\frac{SDpy2.7^2}{Npy2,7} + \frac{SDreg^2}{Nreg}} = 0.63$

The Welch-Satterthwaite degree of freedom Python code (see Appendix F) returns $df = 15.47$

Calculate T-score: $T = \frac{1.47}{0.63} = 2.33$

This T-score, together with the df = 15, correspond with a p-value of 0.034 for a two-tailed hypothesis test, effectively rejecting the null hypothesis.

As a difference is verified, we can now discover whether the difference is negative or positive or both using the confidence interval. By taking the t-value from df_{15} (according to t-value table, we have a t value of 2.13 for a two-tailed test), we can calculate the confidence interval of the difference between two means as follow:

Lower bound = $M_{2.7} - M_{reg} - t*SE = 1.47 - 2.13*0.63 = 0.13$

Upper bound = $M_{2.7} - M_{reg} + t*SE = 1.47 + 2.13*0.63 = 2.81$

Hence we are 95% confident that Py2.7 is more severe than regular strains in terms of fever caused, by resulting in a fever 0.13 to 2.81 degrees F higher than regular strains'.

Similar calculations are also made for other hypotheses. The results are displayed in the following table:

Results \ Hypotheses	Py2.7 vs Regular	Py3 vs Regular	Py2.7 vs Py3
P-value	0.034	0.036	0.74
Reject or Fail to reject null hypothesis	Reject	Reject	Fail to reject
Confidence interval	0.13/2.81	0.09/2.33	-1.36/1.88

3. Results and Conclusion:

It turns out that Py2.7 and Py3 are sever than regular strains. However, we still fail to find any fundamental differences in the fever they cause using statistical methods. Another point to note is that although they are more severe, the highest fever they cause sometimes might not be very different from the fever caused by a regular influenza. (lower bound for Py2.7 and Py3 are 0.13 and 0.09 respectively). Hence, it is highly likely that differentiating them from regular flu using fever temperature is not entirely possible.

APPENDIX B: QUANTIFY THE DIFFERENCES

As shown in the conclusion above, it might be hard sometimes to differentiate between Py2.7 or Py3 and regular fever. Effect size would be the numerical form of this difference as it quantifies the difference between two distributions.

Among the three available effect size measure, we choose to use Hedges' g. Cohen's d assumes that the sample sizes and standard variations between the two samples are approximately equal, which is not very correct when we compare Py2.7 or Py3 with regular strains and Glass' delta might be good for approximation but not optimal when we need precision.

The result of effect sizes are displayed in the table below:

Py2.7 vs Regular	Py3 vs Regular	Py2.7 vs Py3
0.81	0.67	0.11

It could, hence, be safely inferred from the result that Py2.7 and regular strains could be better differentiated than Py3 and regular strains (which are, again, reflected by the confidence interval in appendix A). Py2.7 and Py3, however, have very low effect size, hence low practical significance. This explains the significant resemblance of Py2.7 and Py3 fever distribution in the histograms in Page 2 and the difference in shape of the two new strains and that of regular flu (fatter and longer right tail in the distribution of Py2.7 and Py3). This is also why we recommend our fellow citizens to be concerned about Py2.7 and Py3 as they are not only statistically but also practically difference from the regular flu.

APPENDIX C: STRAIN CHARACTERISTICS - WHITE BLOOD CELL COUNT

While it might be tempting to quantify the characteristics of white blood cell count data using confidence interval, p-value, t-value and effect sizes, such statistical tools are available for normal distributions only. The histograms, and, consequently, the distributions, of this data, however, are multimodal (see page 2). Thus, it would be more sensible to assess this data qualitatively instead of quantitatively considering my available level of knowledge.

The frequency of observations and the range of white blood cell counts are, thus, the two most important characteristics of the histograms. It could noticed from the histogram of Py2.7 that its distribution is bimodal, thus a patient suffering Py2.7 would either have extremely low WBC (<500 WBC) or moderate WBC (7,000-8,000 WBC). It is not sure at which stage of the strain the WBC count is conducted (the 7,000-8,000 WBC count may result from a recovering patient). Nevertheless, it seems certain that abnormally low WBC count is more prevalent in Py2.7 than in regular fever (53.33% of abnormally WBC and 13.64% respectively). The inference is less conclusive in Py3 as the sample size is too small and the observations scatter widely from abnormally low (under 1,700), to low (2,500-3,500), to moderate (5,500-6,500) WBC. This could be an important characteristic that differentiates between Py2.7 and Py3 (Py2.7 patients might be more susceptible to secondary infection). This conclusion (which could be an interesting hypothesis for future research) would not be visible should we consider the WBC distributions normal.

APPENDIX D: CONDITIONAL PROBABILITY FOR DIAGNOSIS

As the danger of Py2.7 and Py3 is known, what is the probability of one citizen being diagnosed of either of these two new strains given that they have an above 102 degree F fever? Bayes theorem might give us a solid answer. The calculations are as follow:

P(new_strains|> 102 degree fever) = $\frac{P(>102\ degree\ fever|new\_strains)*P(new\_strains)}{P(>102\ degree\ fever)}$

According to fever.csv, we have:

P(new_strains|flu) = 39/89 = 43.82%

P(> 102 degree fever|flu) = 29/89 = 32.58%

P(>102 degree fever|new_strains) = 19/39 = 48.72%

Given that a third of Minervopolis is infected. We have:

P(new_strains) = 43.82%*33.33% = 14.61%

P(> 102 degree fever) = 32.58%*33.33% = 10.86%

P(new_strains|> 102 degree fever) = $\frac{48.72\%*14.61\%}{10.86\%}$ = 65.54%

Thus, given that all diagnoses are right (no false negative/ false positive), there is a 65.54% probability that someone having high fever (>102 degree) is infected by either Py2.7 or Py3.

The probability, however, is calculated using the data gathered so far. One common fallacy when consider conditional probability is confusion of the inverse. The result assumes that high fever implies the occurrence of either of the new strains, which may or maynot be true. Our fellow citizens, as a result, should not rely solely on this calculation, which could result in unnecessary paranoia. Other characteristics should be considered also, such as WBC count.

APPENDIX E: DATA ASSUMPTIONS AND THEIR IMPLICATIONS

After conducting multiple calculations, while we could come to a degree of certainty about the danger of Py2.7 and Py3, many assumptions in the process might significantly alternate the result.

The data collected for fever.csv, for example, only takes into account the highest temperature that a patient would experience. Due to the fluctuation of temperature during sickness (and maybe due to medication), regression to the mean, in which an extreme temperature (very severe fever) is followed by a much milder fever, may lead to a false belief of rapid improvement[1]. At the same time, as we also do not know how long the highest fever would last, the data is just a snapshot of a much bigger picture (and it is perfectly acceptable as it is a cross-sectional study).

Thus, on an individual level, taking into account fever temperature to self-diagnose is greatly subject to false belief, confirmation bias and overconfidence. If Py2.7 and Py3 are truly lethal (low white blood cell count implication), paranoid patients could overcrowd medical facilities while optimistic patients would over confidently mistake regression to the mean for the signal of positive improvement, delaying medical intervention until a more dangerous stage, namely secondary infection. Fever should not, as a result, be the sole criteria for diagnosing, research should expand their hypothesis to cover more characteristics for a better understanding of these new strains.

Additionally, while the data for fever.csv is close to a normal distribution, it is highly likely that the real distribution could be right-skewed lognormal due to a natural temperature limit to be defined as a "fever". In such a case, the mean of the data would be larger than the median, and

[1] #regresstomean: Recognize the involvement of regression to the mean in decision making and adjust prediction accordingly.

calculation of confidence interval and hypothesis tests might be misled. It is then necessary that 1/ we determine a proper sample size for conclusive result and 2/ we transform the distribution so that it more closely resemble a normal distribution.

To determine the proper sample size, we refer to Charan and Biswas (2013) as a helpful reference. The formula to calculate the proper sample size for a quantifiable cross-sectional conclusion would, hence, be as follow:

$$N = \frac{Z^2 * SD^2}{d^2}$$

in which N is the desired sample size, Z is the Z-score corresponding to the desired significance level, SD is the standard deviation obtained from the pilot study (which is the SD calculated from fever.csv in this case), and d is the absolute error determined by the researcher.

The proper sample size to determine the mean of the highest fevers caused by Py2.7 at 5% of Type I error, a precision of .5 degrees knowing the standard deviation (from previous studies) of Py2.7 being 2.44 (mentioned in Appendix A), would be:

$$N = \frac{1.96^2 * 2.44^2}{0.5^2} = 91.49$$

Hence, we would need 92 data points instead of 17 to come up with a conclusive inference.

Regarding transformation (to make the distribution closer to normal), a log-transformation might be appropriate because it manipulates bin ranges and reduce outliers (i.e. log(100)-log(10) = log(1000)-log(100) = log(10^4)-log(10^3) =... while 10^4, 10^3, 10^2, and 10 are significantly distant. The lack of precision and sample size in this fever study, however, does not allow further transformation. (or, at least, log transformation)

APPENDIX F: SOURCE CODE

All source codes are written by the author on Python 2.7 using Jupyter notebook.

```
%matplotlib inline

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt


#import data from csv to pandas

wbc = pd.read_csv('./Downloads/WBC.csv')

fever = pd.read_csv('./Downloads/fever.csv')

#pd.DataFrame.hist(fever)

df = pd.DataFrame(fever)


def Histogram(obj, title, x, y, n):

    obj = obj.plot.hist(title = title, ax = axes1[n])

    obj.set_xlabel(x)

    obj.set_ylabel(y)

#program a function to create the first 3 histograms


def Histogram_2(obj, title, x, y, n):

    obj = obj.plot.hist(title = title, ax = axes2[n])

    obj.set_xlabel(x)

    obj.set_ylabel(y)

#program a function to create the second 3 histograms


def Extract (df, param):

    return df[df.Diagnosis.isin([param])]

#extract Py2.7 fever info
```

```
py2 = Extract(df, "Py2.7")

print "Py2.7 Descriptive Stats"

print py2.describe() #descriptive stats for Py2.7

print py2.var()

print ("\n")


#extract Py3 fever info

py3 = df[df.Diagnosis.isin(['Py3'])]

print "Py3 Descriptive Stats"

print py3.describe() #descriptive stats for Py3

print py3.var()

print ("\n")


fig1, axes1 = plt.subplots(nrows = 1, ncols = 3, figsize= (18,5))


Other = Extract(df, "No")

Histogram (Other, 'Regular fever', 'Degrees F', 'Frequency', 0)

print "regular fever descriptive stats"

print Other.describe()

print Other.var()


#plot histogram of Py2.7 fever

Histogram (py2, 'Py2.7 Fever', 'Degrees F', 'Frequency', 1)


#plot histogram of Py3 fever

Histogram(py3, 'Py3 Fever', 'Degrees F', 'Frequency', 2)


fig2, axes2 = plt.subplots(nrows = 1, ncols =3, figsize = (18,5))
```

```
df_wbc = pd.DataFrame(wbc)


Other = Extract (df_wbc, "No")

Histogram_2 (Other, "Regular strain WBC", "WBC", "Frequency", 0)

print "Other WBC\n", Other.describe()


py2_wbc = Extract(df_wbc, 'Py2.7')

Histogram_2 (py2_wbc, 'Py2.7 WBC', 'WBC', 'Frequency', 1)

#print pd.DataFrame.describe(wbc)

print "Py2.7 WBC\n", py2_wbc.describe()


py3_wbc = Extract(df_wbc, 'Py3')

Histogram_2 (py3_wbc, 'Py3 WBC', "WBC", "Frequency", 2)

print "Py3 WBC\n", py3_wbc.describe()
```

```
def Welch(s1,s2,n1,n2):

    x = s1/n1

    y = s2/n2

    welch = (x**2 + y**2)/ ((x**2/(n1-2))+(y**2/(n2-1)))

    return welch

Welch (2.44, 1.54, 17, 50)
```

Reference:

Charan, J., & Biswas, T. (2013). How to Calculate Sample Size for Different Study Designs in Medical Research? *Indian Journal of Psychological Medicine*, *35*(2), 121–126. http://doi.org/10.4103/0253-7176.116232

Effect Size Calculator for T-Test. (n.d.). Retrieved March 16, 2017, from http://www.socscistatistics.com/effectsize/Default3.aspx

Fever in Adults: Learn When to See a Doctor. (n.d.). Retrieved March 16, 2017, from http://www.emedicinehealth.com/fever_in_adults/article_em.htm

Quick P Value from T Score Calculator. (n.d.). Retrieved March 16, 2017, from http://www.socscistatistics.com/pvalues/tdistribution.aspx