Estimating Unobserved COVID-19 Infections in the United States

Quinn White

Submitted to the Department of Statistical and Data Sciences
of Smith College
in partial fulfillment
of the requirements for the degree of
Bachelor of Arts

Ben Baumer, Primary Faculty Advisor
Nicholas Reich, Secondary Faculty Advisor

May 2023

Acknowledgements

Will add

Table of Contents

Chapter 1: Motivation **1**

Chapter 2: Background **3**

2.1 Probabalistic Bias Analysis 5

2.2 Background for the Approach 5

2.2.1 Simple Discrete Example 5

2.2.2 General Solution for the Discrete Case 5

2.2.3 General Solution for the Continuous Case 5

2.2.4 Implementation through the Sampling-Importance-Resampling Algorithm 5

2.3 Bayesian Melding Applied to COVID-19 Misclassification 5

2.3.1 Distribution of $\theta = \{\alpha, \beta, P(S_1|\text{untested})\}$ 5

2.3.2 Direct Prior and Induced Prior Distributions for $P(S_0|\text{test}_+, \text{untested})$ 5

2.3.3 Pooling 5

2.3.4 Derivation of M 5

2.4 Sampling-Importance-Resampling Algorithm 5

2.4.1 Overview 5

2.4.2 Proof that Algorithm Obtains Approximate Sample from Target Distribution 5

2.4.3 Obtaining Logarithmic Pooled Distribution with the Sampling-Importance-Resampling Algorithm 5

2.5 LOESS Smoothing 5

2.5.1 Introduction 5

2.5.2 Fitting the LOESS Curve 5

2.6 Kernel Density Estimation 5

2.6.1 Overview 5

2.6.2 Bounded Density Estimation 5

Chapter 3: Definition of Prior Distributions for Bias Parameters **7**

3.1 Background on the Beta Distribution 8

3.2 Background on the Gamma Distribution 8

3.3 Definition of Prior Distributions for Incomplete Testing Correction . 8

3.3.1 Defining $P(S_1|Untested)$ 8

3.3.2 Defining α 8

3.3.3 Defining β 8

3.3.4 Defining $P(S_0|test+, untested)$ 8
3.4 Definition of Priors for Test Inaccuracy Correction 8
3.4.1 Defining Test Sensitivity (S_e) 8
3.5 Defining Test Specificity (S_p) 8
3.6 Summary Table of Bias Parameter Distributions 8
3.7 Correction for Incomplete Testing 8
3.8 Correction for Diagnostic Test Inaccuracy 8
3.8.1 Derivation of Formula for Correction for Diagnostic Test Inaccuracy 8

Chapter 4: Details of Implementation **9**
4.1 Reproducible Workflow 9

Chapter 5: Results **11**
5.1 County-level 11
5.2 State-level 11
5.3 Comparison to the Covidestim Model 11
5.3.1 Overview 11
5.3.2 The Covidestim Model 11
5.3.3 Assumptions 12
5.4 Comparison to Other Indicators 13
5.5 Limitations of this Comparison 14
5.5.1 Simulation: Bivariate Normal 15
5.5.2 Derivation of the Distribution of X+Y for Bivariate Normal . . 18
5.6 Seropositivity Data 20

Chapter 6: Results **21**
6.1 County-level 21
6.2 State-level 21

Chapter 7: Appendix **23**
7.1 Smoothing Span 23
7.1.1 Changing SPAN for LOESS Smoothing of β 23
7.2 Changing Mean and Variance for Prior Distribution Specifications . . 23

References **25**

Abstract

As we have navigated the COVID-19 pandemic, case counts have been a central source of information for understanding transmission dynamics and the effect of public health interventions. However, because the number of cases we observe is limited by the testing effort in a given location, the case counts presented on local or national dashboards are only a fraction of the true infections. Variations in testing rate by time and location impacts the number of cases that go unobserved, which can cloud our understanding of the true COVID-19 incidence at a given time point and can create biases in downstream analyses. Additionally, the number of cases we observe is impacted by the sensitivity and specificity of the diagnostic test. To quantify the number of true infections given incomplete testing and diagnostic test inaccuracy, this work implements probabilistic bias analysis at a biweekly time scale from January 1, 2021 through February 2022. In doing so, we can estimate a range of possible true infections for a given time interval and location. This approach can be applied at the state level across the United States, as well as in some counties where the needed data are available.

Dedication

You can have a dedication here if you wish.

Chapter 1

Motivation

Placeholder

Chapter 2

Background

Placeholder

2.1 Probabalistic Bias Analysis

2.2 Background for the Approach

2.2.1 Simple Discrete Example

2.2.2 General Solution for the Discrete Case

2.2.3 General Solution for the Continuous Case

2.2.4 Implementation through the Sampling-Importance-Resampling Algorithm

2.3 Bayesian Melding Applied to COVID-19 Misclassification

2.3.1 Distribution of $\theta = \{\alpha, \beta, P(S_1|\textbf{untested})\}$

2.3.2 Direct Prior and Induced Prior Distributions for $P(S_0|\textbf{test}_+, \textbf{untested})$

2.3.3 Pooling

2.3.4 Derivation of M

2.4 Sampling-Importance-Resampling Algorithm

2.4.1 Overview

Example 1:

Example 2:

2.4.2 Proof that Algorithm Obtains Approximate Sample from Target Distribution

2.4.3 Obtaining Logarithmic Pooled Distribution with the Sampling-Importance-Resampling Algorithm

2.5 LOESS Smoothing

2.5.1 Introduction

2.5.2 Fitting the LOESS Curve

2.6 Kernel Density Estimation

2.6.1 Overview

2.6.2 Bounded Density Estimation

Chapter 3

Definition of Prior Distributions for Bias Parameters

Placeholder

3.1 Background on the Beta Distribution

3.2 Background on the Gamma Distribution

3.3 Definition of Prior Distributions for Incomplete Testing Correction

3.3.1 Defining $P(S_1|Untested)$

3.3.2 Defining α

3.3.3 Defining β

3.3.4 Defining $P(S_0|test+, untested)$

3.4 Definition of Priors for Test Inaccuracy Correction

3.4.1 Defining Test Sensitivity (S_e)

3.5 Defining Test Specificity (S_p)

3.6 Summary Table of Bias Parameter Distributions

3.7 Correction for Incomplete Testing

3.8 Correction for Diagnostic Test Inaccuracy

3.8.1 Derivation of Formula for Correction for Diagnostic Test Inaccuracy

Chapter 4

Details of Implementation

- Describe each step here
- Mention reproducible workflow with make

4.1 Reproducible Workflow

Chapter 5

Results

5.1 County-level

5.2 State-level

5.3 Comparison to the Covidestim Model

5.3.1 Overview

One challenge in correcting for biases in general is that although we may have some information about the influence of possible biases, we do not have a ground truth for comparison. However, one approach to handle the fact that the true cases are unobserved is comparing our estimates to those from other approaches seeking to estimate a similar quantity. In particular, if other approaches make different assumptions and come to a similar result, this can give us more confidence in our estimates.

The most notable project seeking to estimate the true infection burden at the county-level over time is the COVIDestim project. In this work, Chitwood et al. proposed a mechanistic model that includes states for asymptomatic/pre-symptomatic infection, symptomatic but mild infection, severe COVID-19 presentations, and death. This approach also enables the estimation of R_t, the number of secondary infections a single infected individual causes at time t.
This is a useful quantity to estimate, but is sensitive to reporting delays and changes in testing practices (https://academic.oup.com/aje/article/190/9/1908/6217341).

5.3.2 The Covidestim Model

Chitwood *et al.* propose a Bayesian evidence synthesis model to correct for reporting delays and time varying case ascertainment testing rate in the estimation of incident infections and R_t.

To estimate the expected cases and deaths at a particular point in time, the model uses a convolution of the time series of observed cases and deaths and reporting delay distributions that are specific to the health state categories. This enables the model to account for the fact that reporting delay is different For any health state, for example, asymptomatic, the individual can either transition to the next health state (symptomatic) or recover. Thus, with each transition between a defined health state, for example, asymptomatic, there is a probability of transitioning to the next health state (in this case, asymptomatic → symptomatic); the complement of this probability is the probability of recovery.

Each of these transitions is defined by a delay distribution. For example, the distribution for moving from asymptomatic to symptomatic represents the probability an individual moves to the symptomatic state at a point in time. The probabilities asymptomatic to symptomatic and symptomatic to severe are modeled as not varying with time. Meanwhile, the probability of transitioning from severe to death was defined to be higher in 2020 due to higher case fatalities early in the pandemic. The infection fatality rates, adjusted to be specific to a given state or county based on age distributions and the prevalence of risk factors for COVID-19, are used to inform the probability of moving from the severe category to the death category.

The change in daily infections from the previous day (i.e., the new infections) is calculated as a function of the estimated effective reproductive number R_t and the mean serial interval, where serial interval is the time from the onset of infection of a primary case to the time of onset of infection in the secondary case. R_t is estimated using a log-transformed cubic spline, under the assumption individuals can only be infected once.

They also defined a distribution for the delay to diagnosis, which was distinct by health state category to reflect differences in diagnosis delays that occur depending on the disease severity. The probability of diagnosis among different health states was allowed to vary by time to reflect changing testing rates throughout the pandemic.

A separate distribution models the reporting delay to correct the total number of diagnoses on a given day for the fact that these diagnoses correspond to past infections.

The observed cases and death data for each state to the model were fitted using negative binomial likelihood functions.

5.3.3 Assumptions

This approach relies on infection fatality ratios and death counts to estimate the true case counts. Thus, it is sensitive to estimates of infection fatality rate, with higher infection fatality ratio estimates resulting in lower estimated infections. The infection fatality ratio is defined as the proportion of COVID-19 infections that lead to death, which means there is uncertainty in estimating both the numerator and the denominator of the ratio. The true cumulative incidence depends on the same uncertainties in estimating the true case burden at any point in time. Estimating the infection fatality ratio itself is a challenging task.

The COVIDestim model uses age-specific estimates of IFR produced by O'Driscoll et al (https://www.nature.com/articles/s41586-020-2918-0). This group used national-level age-stratified, and when possible sex-stratified, COVID-19 death counts and cumulative infection estimates from seroprevalence studies. Of note, the estimates of infection fatality ratio are assumed to be constant over time, which may not be the case due to improving treatments (FIND EXAMPLE) or different variants leading to less severe presentations (FIND PAPER ON OMICRON SEVERITY).

One thing to consider is that infection fatality rate may vary over time, as treatments may vary, as well as the demographics of individuals being infected. For example, during the school year, more students may test positive but will be less likely to die on average than adults (PROVIDE SOURCE FOR THIS). However, these estimates are difficult to acquire; COVIDestim assumed a higher case fatality in 2020 given the novelty of the virus and consequent lack of available treatments.

5.4 Comparison to Other Indicators

There are known issues with seroprevalence estimates. For one, these samples are drawn from a convenience (i.e. nonrandom) sample of individuals with blood specimens taken for purposes other than COVID-19 antibody detection (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html). Secondly, while a positive serological test is evidence for infection, a negative serological test is less clear to interpret. The person may have been infected but not yet have developed antibodies, or their immune system may not have produced antibodies at a detectable level (https://www.cdc.gov/coronavirus/2019-ncov/covid-data/serology-surveillance/index.html).

Indeed, Chitwood et al. found limited concordance between their estimates and seroprevalence data. However, there was a stronger correlation between estimates of cumulative infection and cumulative hospitalizations and cumulative deaths \footnote{ The correlation employed here is the Spearman rank correlation, which measures the strength of the monotonic relationship rather than the strength of the linear relationship, in which case the Pearson correlation coefficient is the usual choice. The Spearman rank correlation is equivalent to the Pearson correlation of the rank values rather than the values themselves (https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient). This distinction is important here since we are interested in the strength of the monotonic relationship rather than the linear relationship between these values. }.

5.5 Limitations of this Comparison

At this point in the pandemic, there is no true gold standard to compare to. Covidestim is one model, among many, that makes key assumptions about aspects of the virus. Another note is that estimates from the Covidestim model are reported on the daily timescale for counties, while the probabilistic approach we implemented here is at the biweekly time scale. For the comparison, we summed the corresponding 2-week intervals for

To ensure the comparisons are on the same time scale, we sum the reported 95% credible intervals for the days in each 2-week interval. These intervals do not represent a 95% credible interval for the 2-week interval, and while such an interval would be ideal for the comparison, computation of a 95% credible interval for the two-week interval is not feasible because of the model structure. Due to the correlation between observations for each day for a given location, summing the intervals yields an estimate that is likely to be more conservative than a true 95% credible interval for the two-week interval would be.

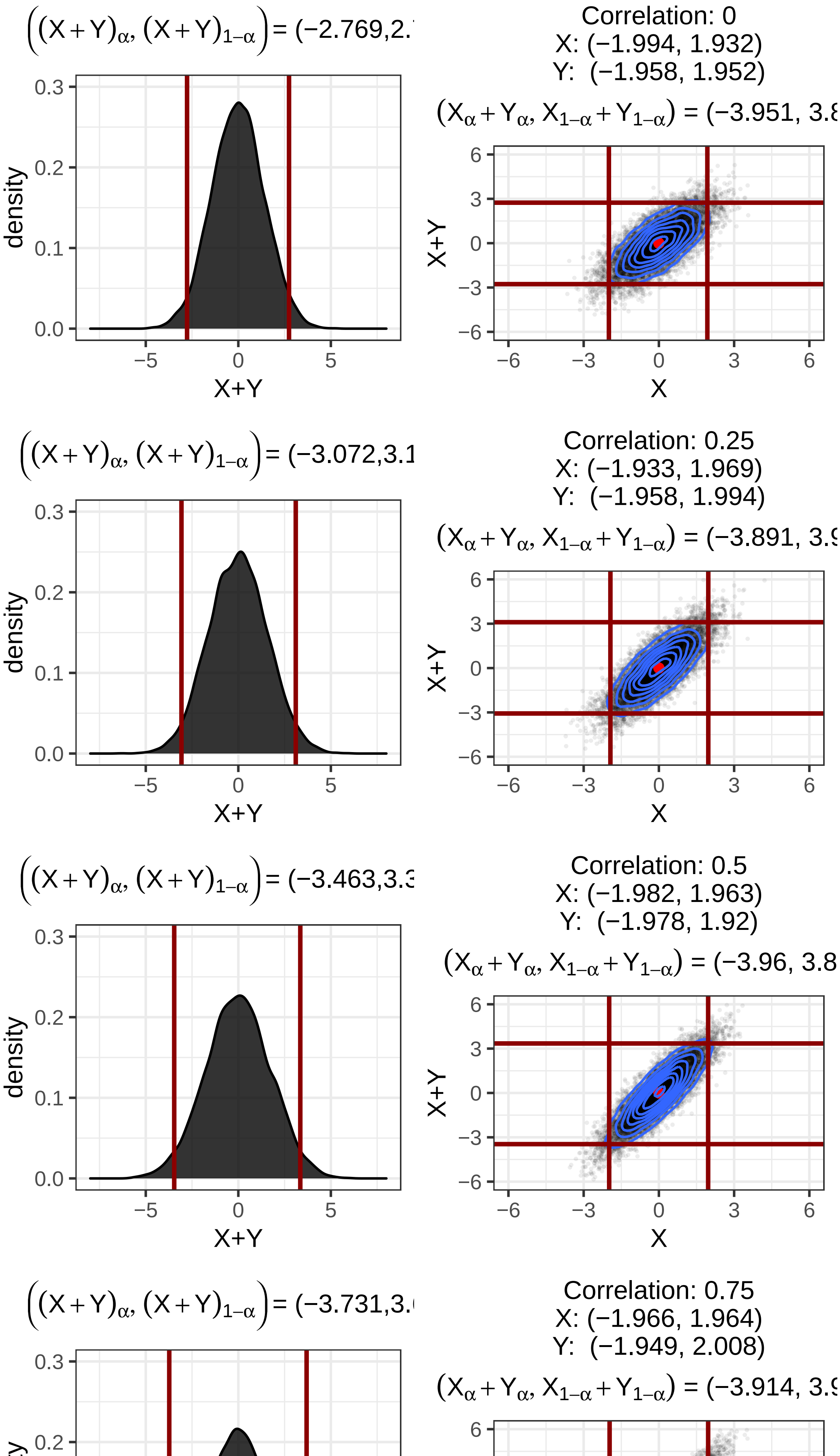
5.5.1 Simulation: Bivariate Normal

We can see this in a concrete example. Let (X, Y) be bivariate normal with $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and correlation matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, and hence where X, Y are marginally standard normal random variables.

We let the subscript α denote the α^{th} and the subscript $1-\alpha$ denote the $(1-\alpha)^{th}$ quantile of the distribution.

In Figure 5.1, in each panel, we increase the correlation ρ between X and Y by 0.25 units and plot the sum $X+Y$ against X. The vertical lines represent quantiles $X_{0.025}$ and $X_{0.975}$, and the horizontal lines represent the quantiles $(X+Y)_{0.025}$ and $(X+Y)_{0.975}$.

We see in Figure 5.1 that when we increase the correlation between X and Y, the width of the interval $\Big((X+Y)_{\alpha}, (X+Y)_{1-\alpha}\Big)$ increases.


$\left((X+Y)_\alpha, (X+Y)_{1-\alpha}\right) = (-2.769, 2.7$
0.3
0.2
0.1
0.0
density
−5
0
5
X+Y
Correlation: 0
X: (−1.994, 1.932)
Y: (−1.958, 1.952)
$(X_\alpha + Y_\alpha, X_{1-\alpha} + Y_{1-\alpha}) = (-3.951, 3.8$
6
3
0
−3
−6
X+Y
−6
−3
0
3
6
X
$\left((X+Y)_\alpha, (X+Y)_{1-\alpha}\right) = (-3.072, 3.1$
0.3
0.2
0.1
0.0
density
−5
0
5
X+Y
Correlation: 0.25
X: (−1.933, 1.969)
Y: (−1.958, 1.994)
$(X_\alpha + Y_\alpha, X_{1-\alpha} + Y_{1-\alpha}) = (-3.891, 3.9$
6
3
0
−3
−6
X+Y
−6
−3
0
3
6
X
$\left((X+Y)_\alpha, (X+Y)_{1-\alpha}\right) = (-3.463, 3.3$
0.3
0.2
0.1
0.0
density
−5
0
5
X+Y
Correlation: 0.5
X: (−1.982, 1.963)
Y: (−1.978, 1.92)
$(X_\alpha + Y_\alpha, X_{1-\alpha} + Y_{1-\alpha}) = (-3.96, 3.8$
6
3
0
−3
−6
X+Y
−6
−3
0
3
6
X
$\left((X+Y)_\alpha, (X+Y)_{1-\alpha}\right) = (-3.731, 3.$
0.3
0.2
density
Correlation: 0.75
X: (−1.966, 1.964)
Y: (−1.949, 2.008)
$(X_\alpha + Y_\alpha, X_{1-\alpha} + Y_{1-\alpha}) = (-3.914, 3.9$
6

In Figure 5.2, we compare the intervals defined by taking the quantiles of the sum, $\Big((X+Y)_\alpha, (X+Y)_{1-\alpha}\Big)$, to the intervals taken by summing the quantiles individually, $\Big(X_\alpha + Y_\alpha,\ X_{1-\alpha}+Y_{1-\alpha}\Big)$. We notice that, as we saw in Figure 5.1, increasing the correlation increases the width of the interval $\Big((X+Y)_\alpha, (X+Y)_{1-\alpha}\Big)$, while the interval $\Big(X_\alpha+Y_\alpha,\ X_{1-\alpha}+Y_{1-\alpha}\Big)$ is constant since changing the correlation does not change the marginal quantiles $X_\alpha, X_{1-\alpha}$.,

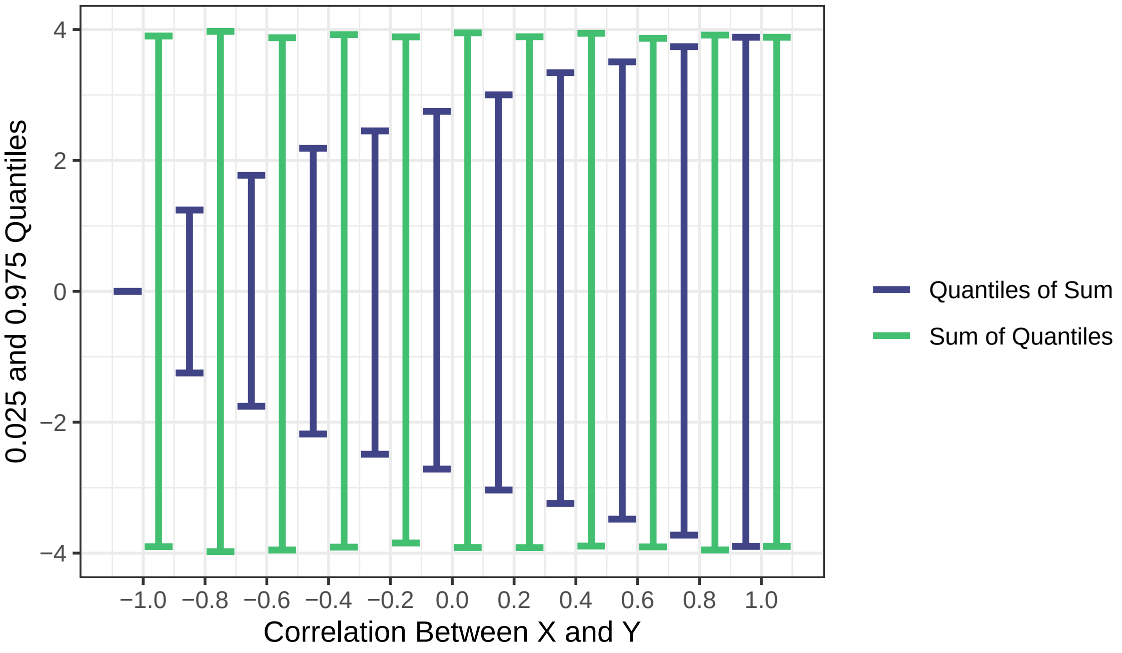


Figure 5.2

As we see in Figure 5.2, the intervals are identical when X, Y are perfectly correlated. This result is not dependent on the choice of distribution, as we can show by considering CDFs and quantile functions of a general distribution.

Quantiles of the Sum of Perfectly Correlated Random Variables

When two random variables X and Y are perfectly correlated,

$$X_\alpha + Y_\alpha = (X+Y)_\alpha.$$

When X and Y are perfectly correlated, Y must be a linear combination of X, so we can write $X+Y = X+bX = (1+b)X$.

Then, let the α^{th} quantile of $(1+b)X$ be x_α. By definition of the quantile function, we have

$$F^{-1}_{(1+b)X}(\alpha) = x_\alpha \implies P((1+b)X \leq x_\alpha) = \alpha.$$

Since $(1+b)$ is just a constant, we can divide to yield

$$P\Big(X \leq x_\alpha/(1+b)\Big) = \alpha.$$

To optain hte quantile for bX, we can multiply each side by b to yield

$$P\Big(bX \leq bx_\alpha/(1+b)\Big) = \alpha.$$

Putting these results together, we have

$$F_{bX}^{-1}(\alpha) + F_X^{-1}(\alpha) = \frac{bx_\alpha}{1+b} + \frac{x_\alpha}{1+b} = x_\alpha \qquad\qquad = F_{(1+b)X}^{-1}(\alpha)$$

5.5.2 Derivation of the Distribution of X+Y for Bivariate Normal

We can see why we observe this relationship between intervals based on the the sum of the α^{th} quantiles of the individual distributions, $X_\alpha + Y_\alpha$, and the intervals based on the α^{th} quantile of the distribution of $X+Y$ by considering the definition of the quantile function of the normal distribution.

Defining $Z = g(X,Y) = X+Y$, we can obtain the density function by a change of variables. Notice if $g(X,Y) = X+Y$, $g^{-1}(X,Z) = Z - X$, so we have

$$\begin{aligned}
f_{X,Z}(x,z) &= f_{X,Y}(x, g^{-1}(x,z)) \left|\frac{\partial g^{-1}(x,z)}{\partial z}\right| \\
f_{X,Z}(x,z) &= f_{X,Y}(x, z-x) \left|\frac{\partial (x-z)}{\partial z}\right| \\
f_{X,Z}(x,z) &= f_{X,Y}(x, z-x)\,|1| \\
f_{X,Z}(x,z) &= f_{X,Y}(x, z-x)
\end{aligned}$$

Then, we can marginalize out X to get the PDF of f_Z by taking

$$f_Z(z) = \int_{\infty}^{\infty} f(x, z-x)\, dx.$$

Since (X,Y) is bivariate normal with correlation ρ, the PDF is given by

$$f(x,y) = \frac{exp\left[\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_x^2} - \frac{2\rho(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right)\right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

Integrating with respect to x[1], we have

[1] This integration is extremely long and technical, so we do not include it here.

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{\exp\left[\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_x^2} - \frac{2\rho(x-\bar{x})(z-x-\bar{y})}{\sigma_x\sigma_y}\right)\right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}dx$$

$$= \frac{\exp\left[-\frac{(z-(\bar{x}+\bar{y}))^2}{2(\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y)}\right]}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y)}}.$$

It follows that Z is a normal random variable with mean $\bar{x}+\bar{y}$ and standard deviation $\sqrt{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}$.

In Figure 5.3, we plot the density estimate of the distribution of $X+Y$ for $(X,Y) \sim MVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.2\\0.2 & 1\end{pmatrix}\right)$ and plot the density of the random variable $X+Y = Z \sim N\left(\bar{x}+\bar{y}, \sqrt{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}\right)$ and see they are in close alignment, as expected.

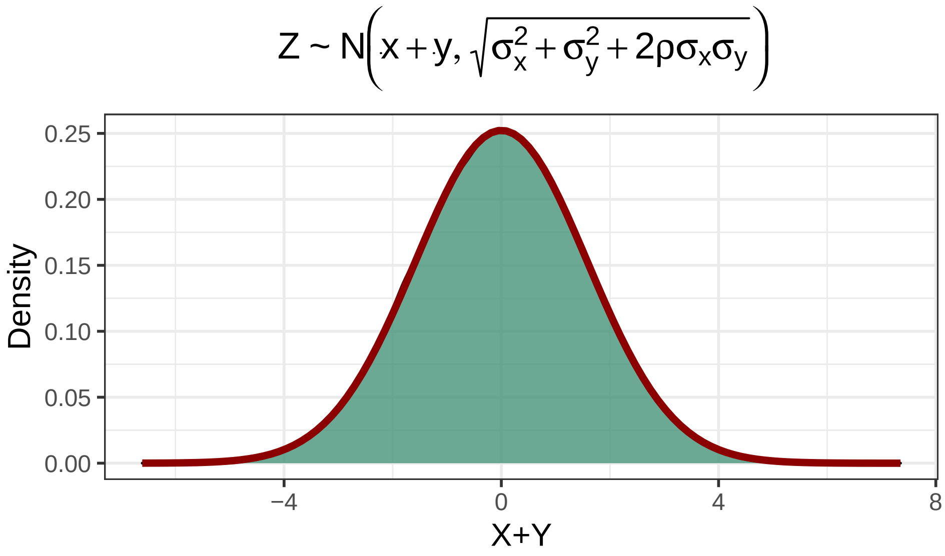


Figure 5.3: The theoretical density of $N\left(\bar{x}+\bar{y}, \sqrt{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}\right)$ is plotted in red over the kernel density estimate of the observed distribution of $X+Y$.

Since we now know $Z \sim N\left(\bar{x}+\bar{y}, \sqrt{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}\right)$, we can consider the quantile function of the normal distribution, which is defined as

$$F_Z^{-1}(\alpha) = \mu + \sigma_Z \operatorname{erf}^{-1}(2\alpha - 1).$$

and since $\sigma_Z = \sqrt{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}$ we have

$$F_Z^{-1}(\alpha) = \mu + \left(\sqrt{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}\right) \operatorname{erf}^{-1}(2\alpha - 1).$$

Now, we note the inverse error function erf^{-1} is increasing (Figure 5.4).

This means if $\alpha > 0.5$, F_Z^{-1} is increasing with increasing values of ρ, and if $\alpha < 0.5$, F_Z^{-1} is decreasing with increasing values of ρ.

This if we have a pair of correlated random variables (X_1, Y_1) and (X_2, Y_2) and $\rho_{X_1,Y_1} > \rho_{X_2,Y_2}$ and consider $\alpha < 0.5$,

$$(X_1 + Y_1)_\alpha < (X_2 + Y_2)_\alpha$$

and

$$(X_1 + Y_1)_{1-\alpha} > (X_2 + Y_2)_{1-\alpha}.$$

This is exactly what we observed in Figure 5.2.

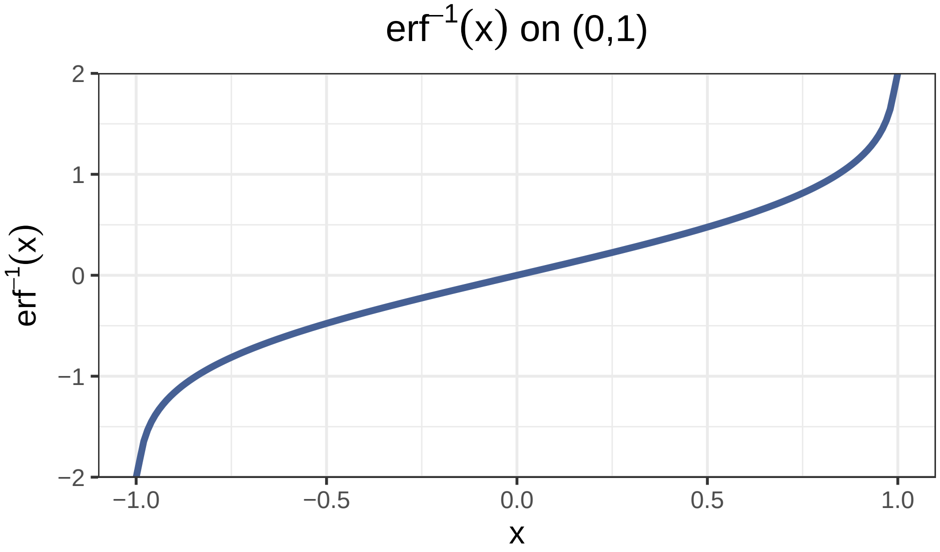


Figure 5.4

5.6 Seropositivity Data

To add

Chapter 6

Results

6.1 County-level

6.2 State-level

Chapter 7

Appendix

Placeholder

7.1 Smoothing Span

7.1.1 Changing SPAN for LOESS Smoothing of β

7.2 Changing Mean and Variance for Prior Distribution Specifications

References

Placeholder