PROGRAMA DE PÓS -GRADUAÇÃO EM ORGANIZAÇÕES E MERCADOS – MESTRADO EM ECONOMIA APLICADA

PPGOM

UFPEL

WORKING PAPER

Estimação dos níveis de infecção por covid19 no brasil

01/2020

OUTUBRO

Erik Figueiredo (PPGE-UFPB)
Démerson André Polli (UnB)
Bernardo Borba de Andrade (UnB)

RELATÓRIO DE PESQUISA:

ESTIMAÇÃO DOS NÍVEIS DE INFECÇÃO POR COVID19 NO BRASIL

Erik Figueiredo

eafigueiredo@gmail.com

Universidade Federal da Paraíba

Démerson André Polli

polli@unb.br

Universidade de Brasília

Bernardo Borba de Andrade

bbandrade@unb.br

Universidade de Brasília

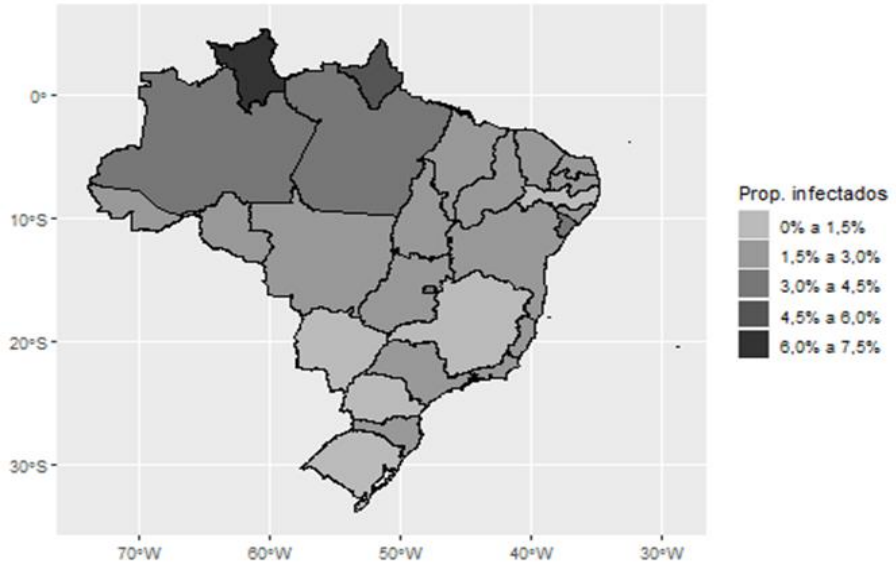
Resumo: Os dados da PNAD-COVID19 indicam que, pelo menos, de 2% da população brasileira já foi infectada pelo COVID19 até agosto de 2020. Postula-se que o real número de infectados é bem maior do que o revelado por testagem pois (i) há um número limitado de testes, o que implica que muitos dos infectados, em especial, aqueles assintomáticos ou com sintomas leves, não realizam exames e (ii) os testes reportados na pesquisa – em especial, os testes "rápidos" – tendem a apresentar um número elevado de resultados falsos-negativos. Usando um método capaz de corrigir os vieses decorrentes de (i) e (ii), calculou-se que os números brasileiros podem estar subestimados em praticamente cinco vezes, número conservador se forem considerados valores obtidos por algumas modelagens epidemiológicas de notoriedade. O método utilizado foi proposto por Wu et al. (2020) e se baseia em metodologia semi-bayesiana. Aplicando-se o método, com pequenas variações, aos dados brasileiros obtidos da PNAD COVID-19, estima-se uma taxa de infecção nacional em torno de 10,8%, podendo chegar a 16,6%. Observou-se elevada heterogeneidade estadual, com níveis de infecção variando entre 5,8% (Rio Grande do Sul) e 30% (Roraima). O fato de que nem toda a população está sujeita ao contágio, pois, alguns indivíduos podem apresentar imunidade natural ou cruzada deve ser somado a esta porcentagem para se obter a porcentagem de imunizados, mas este dado ainda é difícil de mensurar. A estimação, a partir dos casos confirmados, deve ser discutida à luz do debate relativo ao ponto da imunidade coletiva. Alguns estudos sugerem que as estimativas padrão superestimam esse ponto e, quando calculado de forma apropriada, o limiar se situaria entre 10% e 20% nos países europeus. Portanto, a depender da real taxa de infecção, de fatores ligados à imunidade pré-existente e ao limiar da imunidade coletiva, há evidências de que alguns estados brasileiros estejam próximos ou já tenham alcançado a imunização coletiva necessária para a redução da contaminação a nível controlável.

Palavras-chave: PNAD-COVID19; Análise Quantitativa de Viés; Estimação Semi-Bayesiana; Imunidade Coletiva

1. INTRODUÇÃO

A divulgação da Pesquisa Nacional por Amostra de Domicílios – PNAD COVID19 (IBGE, 2020), trouxe, como destaque inicial, o número de pessoas submetidas aos testes para a COVID19 até agosto de 2020. Cerca de 21,6% dos que realizaram testes, quase 2% da população total, obtiveram resultados positivos. A utilização desse banco de dados permite avaliar a distribuição desses números em nível estadual. A Figura 1 ilustra a distribuição geográfica do percentual de testes positivos em relação à população total.

Figura 1 - Proporção de casos confirmados de COVID19 por estado em relação à população total.



Fonte – Dados da PNAD-COVID19, Agosto /2020.

Observa-se uma elevada heterogeneidade regional, com destaques para o número de infectados nos Estados do Norte do país, em especial Roraima e Amapá, com 7,4% e 5,9% da população infectada, respectivamente (ver Tabela A3 em anexo). O Distrito Federal também merece destaque, com 4,8%. No Nordeste, o maior número de infectados é observado em Sergipe (3,4%), seguido por Piauí (2,6%) e Maranhão (2,6%). Já as menores taxas de infecção são registradas no Rio Grande do Sul e Minas Gerais, ambas em torno de 0,9% (as únicas do país abaixo de 1%).

Os números acima servem como um importante guia para as políticas públicas de saúde e determinação de retomada das atividades socioeconômicas. Contudo, as contagens baseadas somente em indivíduos submetidos a teste não captam o número total de infectados. As razões para isso são conhecidas e bastante discutidas na literatura especializada: (i) há um número limitado de testes e diferentes padrões de acessibilidade; (ii) muitos dos infectados, em especial aqueles assintomáticos ou com sintomas leves, não realizam exames, mesmo com testes disponíveis (Pearce, 2020); (iii) os testes reportados na PNAD-COVID19 incluem os testes "rápidos" e outros tipos que tendem a apresentar um número elevado de resultados falsos-negativos (ver, entre outros, Lan et al., 2020 e Yang et al., 2020) e, por fim, (iv) a projeção do percentual de resultados positivos para a população deve considerar o viés de seleção contido nesse tipo de abordagem, qual seja, o indivíduo busca realizar o teste, caso tenha suspeita de estar doente. Uma discussão geral sobre viés de seleção pode ser encontrada em Angrist and Pischke (2009).

Esse padrão tende a ser reforçado por medidas públicas (hoje já evitadas) de desincentivar a busca por atendimento precoce. Em meados de abril, o Núcleo de Operações e Inteligência em Saúde (NOIS), que reúne colaboradores da PUC-RJ, Fiocruz e Instituto D'Or, estimou que o número de infectados no Brasil seria até 12 vezes maior do que o reportado, valor próximo à média de 10,5 estimada para diversos países (Rahmandad et al., 2020). A Figura A2 no Anexo ilustra levantamento feito pela revista *The Economist* e traz a razão entre estimativas e infectados a partir de inquéritos sorológicos e casos confirmados para um grupo de países e cidades selecionados. Para os Estados Unidos, por exemplo, tem-se uma estimativa de que o número de infectados seja 7 vezes o número de casos confirmados. Este valor é compatível com estudo recente (Wu et al., 2020), que detalharemos a seguir, o qual estima em 9 vezes o fator de subnotificação com dados de abril para os Estados Unidos.

Dado a diferença entre casos confirmados e número de infectados, como podemos estimar o número de infectados? Uma primeira abordagem é recorrer a análises quantitativas simples e comparativos a partir de dados históricos de doenças respiratórias e, possivelmente, de outras informações epidemiológicas e demográficas. Por exemplo, em maio deste ano, Ribeiro e Bernardes (2020) utilizaram estatísticas de hospitalizações por Síndrome Respiratória Aguda Grave (SRAG) levantadas pela Fiocruz para estimar o real número de hospitalizações por COVID19. O estudo concluiu que casos mais graves seriam 3,8 vezes o número notificado e que, considerando-se os casos que não geram hospitalização, o índice de subnotificação seja ainda maior.

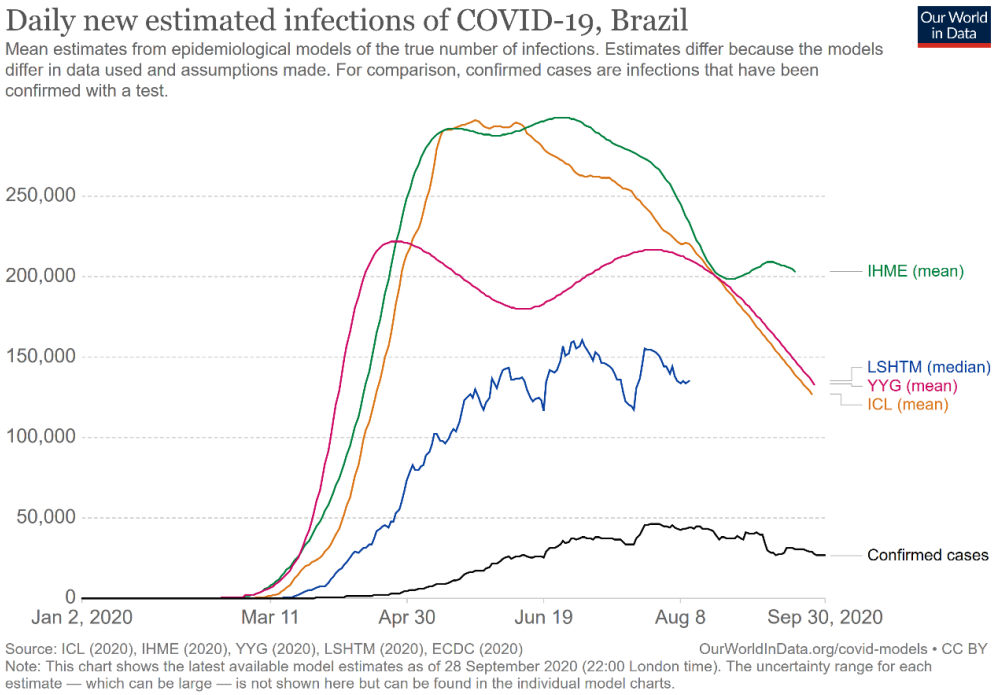
Uma outra abordagem para se estimar o número de infectados é utilizar modelos epidemiológicos, como por exemplo, os modelos compartimentais SEIR (Suscetível-Exposto-Infectado-Recuperado). Há uma gama extensa de variações desses modelos, os

quais tentam reproduzir matematicamente a dinâmica de uma dada epidemia. Alguns desses modelos têm sido alvo de críticas por produzirem números elevadíssimos de infectados em períodos relativamente curtos de tempo. É grande o desafio de se modelar formalmente os diversos elementos atuantes na transmissão e imunização de um agente patogênico numa dada população (com possíveis fatores migratórios e intervenções). Além disso, tem-se a dificuldade de se calibrar os muitos parâmetros com informações clínicas e epidemiológicas imprecisas e poucos dados no caso de uma agente infeccioso novo. Ainda assim, alguns desses modelos têm tido ampla divulgação e citação. Se considerarmos as estimativas mais recentes de infecções por SARS-CoV-2 oriundas de alguns destes modelos (Figura 2), o número de infectados seria muitas vezes o número de casos confirmados. Por exemplo, os relatórios públicos mais recentes do modelo SEIR implementado por Youyang Gu[1] (YYG) indica que o número de infectados já teria alcançado 16,9% da população no Brasil (estimativa de 30/09/2020), ou seja, quase 8 vezes a porcentagem atual (2,2%) de casos confirmados. Já o modelo da *London School of Hygiene & Tropical Medicine*[2] (LSHTM) estima que de 28% a 42% dos casos **sintomáticos** não estariam sendo notificados. Os modelos do *Imperial College London* (ICL) e do *Institute for Health Metrics and Evaluation* (IHME) já apontam para número de casos muito maiores do que os modelos YYG e LSHTM (cf. Figura 2).

Uma terceira abordagem, baseada na técnica estatística de **análise quantitativa de viés** (Lash et al. 2011), foi explorada em artigo recém-publicado por Wu et al. (2020). Em resumo, os autores utilizam uma estratégia de correção probabilística de viés semi-Bayesiana, demonstrando que os números estaduais de casos positivos reportados no Estados Unidos precisam ser corrigidos em 3 a 20 vezes: "*Accounting for uncertainty, the number of infections during this period was 3 to 20 times higher than the number of confirmed cases*." (Wu et al., 2020, p. 2). A metodologia utilizada é fortemente calcada em dados de testagem fazendo correções para a (baixa) abrangência da testagem e para imprecisões nos testes. Os pesquisadores utilizaram evidência da literatura científica para desenhar modelos probabilísticos (distribuições *a priori*) adequados para as quantidades que afetam diretamente o número esperado de infectados a partir do número de casos confirmados. Trata-se de um trabalho inovador e bem documentado que iremos reproduzir para os estados brasileiros e contrastar seus resultados com outras estimativas amplamente divulgadas.

[1] https://covid19-projections.com/ (acessado em 30/09/2020)
[2] https://cmmid.github.io/topics/covid19/global_cfr_estimates.html (acessado em 21/09/2020)

Figura 2 - Evolução do número de casos de COVID19 no Brasil segundo diferentes modelos epidemiológicos e por registro de casos confirmados para 28/09/2020.



Fonte – Reproduzido de OurWorldInData.org (arquivo png disponível para uso público).

Na próxima seção apresentamos estimativas para o viés de subnotificação utilizando a metodologia de Wu et al. (2020). A seção 3 discute questões de imunidade de rebanho, com foco no número a partir do qual a taxa de infecção começa a cair mais fortemente e a transmissão se mantém endêmica ou nula. A seção 4 reúne as considerações finais. Aspectos metodológicos e de estimação serão reportados no Anexo 1. O Anexo 2 é destinado a resultados adicionais, incluindo uma estimativa de eliminação do viés de seleção via criação de uma sub-amostra da PNAD-COVID19: trata-se de uma estimativa mais simples do que a obtida por análise probabilística de viés mas de fácil implementação e que nos serve de modo comparativo. Diante de uma situação nova e com elevada incerteza faz-se mister a maior utilização possível dos dados disponíveis.
Ressaltamos, por fim, a importância da PNAD-COVID19. Trata-se de pesquisa de grande relevância que produz uma base de dados fruto de planejamento amostral, em nível nacional, diferentemente do que ocorre com os dados para os EUA utilizados por Wu et al. (2020) que tiveram que ser compilados de diferentes bases estaduais.

2. MÉTODOS

Para quantificar as subnotificações dos casos da COVID19 aplicamos o método de correção de viés desenvolvido por Wu et al. (2020) utilizando os dados da PNAD COVID19 de agosto/2020.

A PNAD COVID19 tem por objetivo mensurar o impacto da pandemia de COVID19 no mercado de trabalho brasileiro. A pesquisa teve início em 04 de maio de 2020, com entrevistas realizadas por telefone em, aproximadamente, 48 mil domicílios por semana, com um total de 193 mil domicílios por mês, em todo o país. Os domicílios entrevistados no primeiro mês de coleta de dados permanecem na amostra nos meses subsequentes, até o término da pesquisa. Para a confecção deste artigo utilizamos os dados coletados no mês de agosto de 2020.

A análise quantitativa de viés (Lash et al., 1999) permite corrigir vieses em estimativas oriundas de estudos observacionais a partir de suposições relativas à direção desses vieses (como viés de seleção, viés de informação/classificação, fatores de confundimento não mensurados, etc.). Métodos probabilísticos de correção de viés são mais flexíveis e permitem a correção múltipla de vieses. Os métodos probabilísticos com estrutura Bayesiana (em geral Bayesiana empírica ou semi-Bayesiana) são ainda mais flexíveis mas com custo computacional adicional e necessidade de programação específica para cada caso. O método é denominado semi-Bayesiano porque define a *priori* distribuições para os parâmetros que afetam o viés, mas sem utilizar a função de verossimilhança para conectar a distribuição dos parâmetros aos dados e produzir estimativas dos parâmetros. O objetivo da modelagem proposta por Wu et al. (2020) é obter estimativas mais acuradas diante de testagens incompletas ou com elevada chance de reportar resultados falso-negativos. Um resumo da metodologia utilizada por Wu et al. (2020) e adotada aqui se encontra no Anexo 1.

Em nossa formulação, mantivemos as formas funcionais desenvolvidas por Wu et al. (2020) mas foi necessário ajustar os parâmetros para valores condizentes com o observado no Brasil (ver Tabela A1 do Anexo 1). O método se mostrou robusto a variações razoáveis das especificações que fizemos.

Vamos reportar os resultados em termos da porcentagem estimada de infecções e do "fator de subnotificação" (FS), tal que

$$\text{Número de Infectados} = FS \times \text{Casos Confirmados}.$$

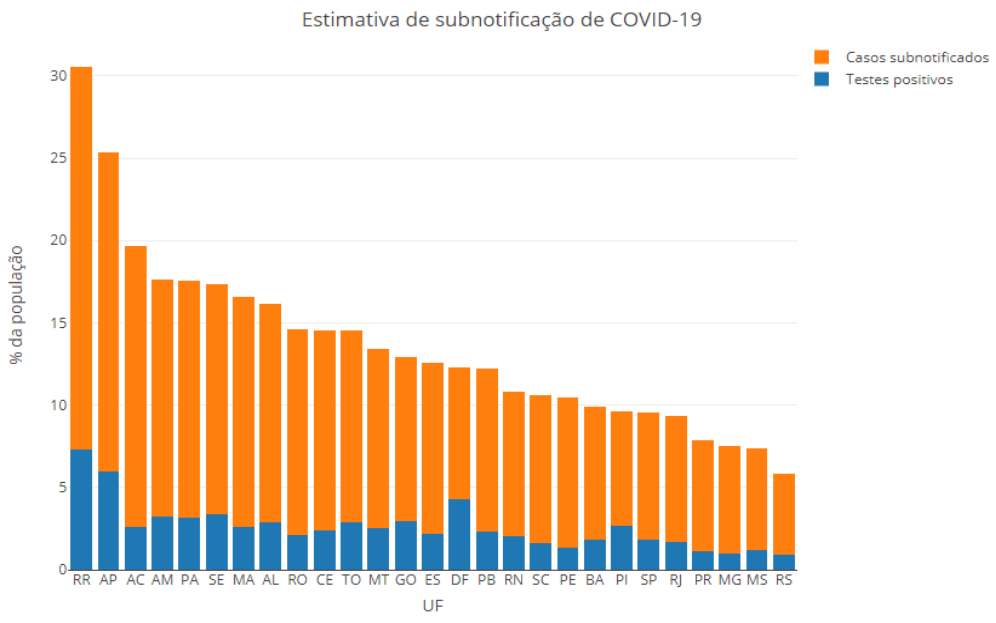
A metodologia aqui utilizada difere substancialmente dos modelos compartimentais mencionados anteriormente e é estritamente baseada em evidência

estatística, sem pretensão de modelar mecanismos de transmissão ou traçar a dinâmica do contágio.

3. RESULTADOS DA ANÁLISE PROBABILÍSTICA DE VIÉS PARA O BRASIL

Os principais resultados estão reportados na Figura 3. Em síntese, são reunidas duas informações para cada estado brasileiro: a barra **azul** representa o percentual de infectados observado nos dados da PNAD-COVID19 e a parte **laranja** da barra representa os valores recuperados pelo modelo probabilístico. Portanto, a soma dessas duas áreas indica o número estimado de infectados. Destaca-se o comportamento dos Estados da região Norte. Por exemplo, em Roraima, o número reportado de casos confirmados na PNAD-COVID19 é de 7,4%. Quando aplicada a correção de viés, esse o número estimado de infectados é de 30%. Observamos ainda que para o estado do Amazonas tem-se a estimativa de 17,6% de infectados e que estudo recente (Buss et al., preprint) (ainda não revisado por pares) sugere que 66% da população de Manaus já possa ter sido infectada. O número, ainda que surpreendentemente alto e que venha a sofrer revisão para baixo, corrobora a noção geral de que há forte subestimação do número de infectados se olharmos somente para os casos confirmados.

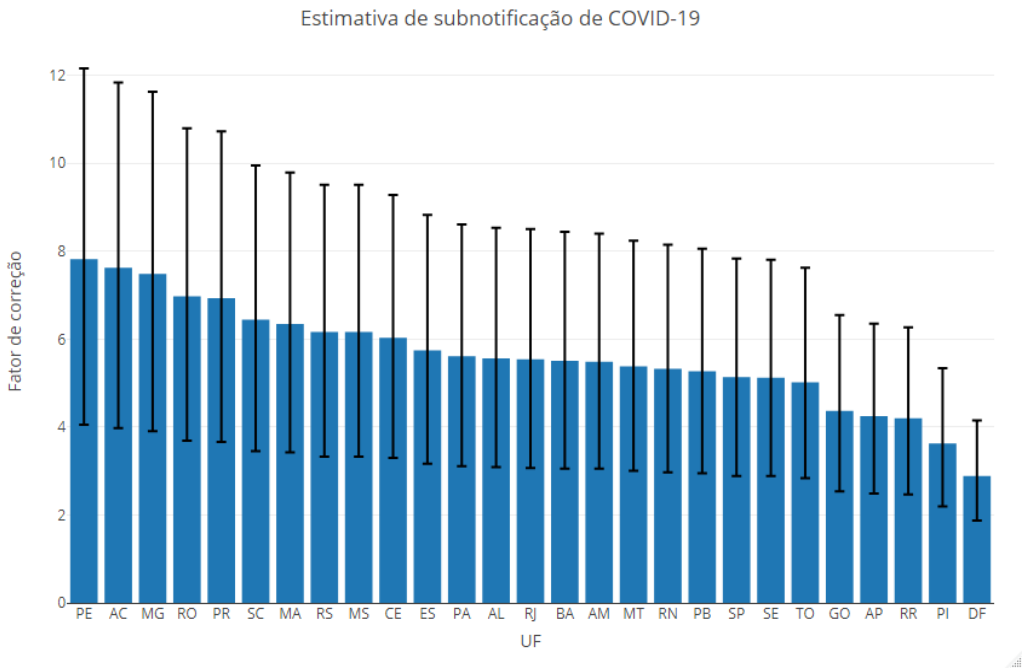
Figura 3: Casos confirmados e estimativa do número de infectados de COVID19 por estado.



Fonte – PNAD-COVID19.

A Figura 4 (ver ainda Tabela A3) ilustra a magnitude da subnotificação dos casos de COVID19, incluindo os limites inferior e superior (intervalo de credibilidade de 95%) para as estimativas. A barra indica em quantas vezes o número reportado oficialmente precisa ser inflado para se chegar ao número real de infecções. Nesse cenário destaca-se Pernambuco. Em média o valor reportado oficialmente está 8 vezes menor do que o valor corrigido pelo viés. Os limites inferior e superior indicam que esse número poderia estar 4 ou 12 vezes maior.

Figura 4: Estimativa do fator de subnotificação de casos confirmados de COVID19 por estado



Fonte – PNAD-COVID19.

Como essas estimativas se comparam a valores obtidos em outros estudos? Até o momento, o Brasil registra número de casos confirmados que corresponde a 2,2% da população, próximo ao valor de 1,9% obtido a partir dos resultados mais recentes da PNAD-COVID19. No entanto, pelos diversos motivos conforme mencionados na seção anterior, estes números nos dão apenas uma indicação da real proporção de infectados. Pelos resultados coletados de outros estudos e nossas estimativas (Figura 4 e Tabela A3), temos indicação de que a proporção de infectados no Brasil pode estar entre 3,7% (de modo bastante conservador) e algo próximo de 20%:

- De acordo com o estudo do NOIS e valores internacionais estimados por Rahmandad (2020) o número de infectados seria em torno de 10 vezes mais do que o reportado e, portanto, o Brasil teria hoje **20%** ou mais de infectados. Os

modelos epidemiológicos ICL e IHME sugerem valores comparáveis.

- O modelo epidemiológico YYG estima a proporção de infectados no Brasil em **16,9%**.

- Nossas estimativas com base na metodologia de Wu et al. (2020) é de que a proporção de infectados no Brasil estava, no fim de agosto, em **10,8%**.

- Segundo Ribeiro e Bernardes (2020) os casos graves seriam 3,8 vezes o reportado. Adotando este fator para todos os casos temos uma estimativa conservadora (dado que a subnotificação entre casos não graves deve ser muito maior) de **8,4%** da população já infectada no Brasil.

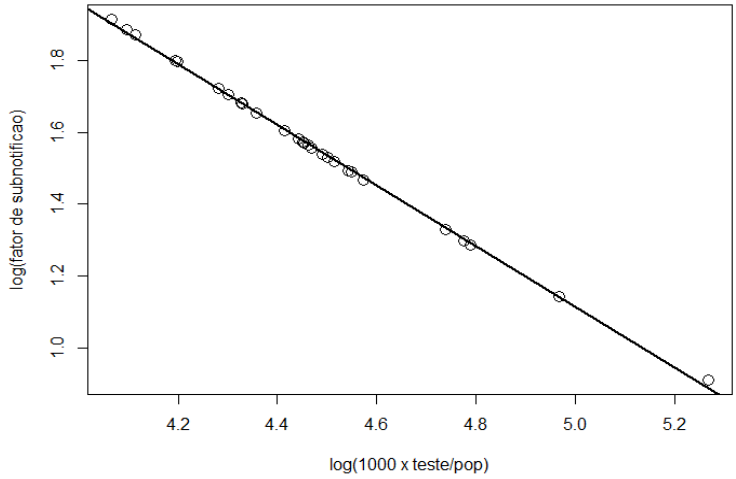
- O modelo LSHTM estima que os casos sintomáticos reportados no Brasil são apenas 65% dos casos sintomáticos com intervalo de 95% de confiança entre 58% e 72%. De acordo com a PNAD COVID-19, 22% dos entrevistados com resultado positivo (que por sua vez são 1,84% da população) apresentaram algum sintoma. Uma possível estratégia é adotar a estimativa de 65% para os sintomáticos e o valor inferior de 58% para os assintomáticos (dentre os quais esperamos maior subnotificação). Dessa forma, estimaríamos a proporção de infectados no Brasil de modo bastante conservador em **3,7%**.

Os resultados da análise de viés sugerem uma regra simples de obtenção do fator de subnotificação a partir da alta correlação linear que se verifica entre o log do fator de subnotificação e o log da razão de testagem (Figura 5),

$$\mathrm{FS} \approx 0.6076 \left(\frac{\text{testes realizados}}{\text{populacao}}\right)^{-0.845}.$$

Dessa forma, se quisermos estimar o FS para uma dada região que realizou testes em 5% a 20% da população (valores contidos na nossa amostra) teremos os valores da Tabela 1.

Figura 5 – Relação entre as estimativas do fator de subnotificação e a razão de testes realizados (por 1000 habitantes) em escala log-log para 27 UFs.



Fonte – PNAD COVID19.

Tabela 1 – Valores aproximados para o fator de subnotificação para diferentes percentuais de testagem na população assumindo as estimativas com base na PNAD Covid19 de agosto/2020.

Porcentagem testada	**5%**	**10%**	**15%**	**20%**
Fator de subnotificação	7,6	4,3	3,0	2,4

4. CONSIDERAÇÕES ADICIONAIS

4.1. LIMIAR DA IMUNIDADE DE REBANHO

Uma vez estimado o número de infectados, surgem, naturalmente, questionamentos sobre a imunidade coletiva (ou imunidade de rebanho). A imunidade coletiva é uma possibilidade para o enfraquecimento ou eventual desaparecimento de uma doença infecciosa. Em geral, se refere à porcentagem da população que, uma vez imunizada, faz com que a transmissão do agente patogênico se limite à forma endêmica[3].

[3] Mais precisamente, a porcentagem em si é denominada de *limiar da imunidade de rebanho* (HIT na sigla usual, em inglês).

Diante disso, questiona-se se é possível atingir imunidade coletiva para a COVID-19 e, em caso afirmativo, qual o limiar a ser alcançado.

A possibilidade de que a imunidade coletiva possa não ser aplicável ao novo coronavírus, SARS-CoV-2, pode ser argumentada, por exemplo, com base na epidemiologia do coronavírus HCoV-NL63 (Kiyuka et al., 2018) sob argumentos de que não há comprovação de imunização por período de tempo longo o suficiente e que as reinfecções seriam frequentes e até mesmo mais intensas nas recorrências. No entanto, nosso entendimento da literatura epidemiológica até o momento (por exemplo, Gudbjartsson et al., 2020) e de opiniões especializadas tornadas públicas é de que a imunidade de rebanho é alcançável para o novo coronavírus, SARS-CoV-2, e iremos trabalhar com esta hipótese. Dessa forma, faz-se necessário saber qual o limiar a ser alcançado.

A fórmula padrão para o cálculo do limiar de imunidade do rebanho é $1 - 1/R_0$, sendo R_0 o número de reprodução básico, que é o número de pessoas, em média, que seria infectado por cada pessoa que adoece, assumindo que 100% das pessoas na população sejam igualmente suscetíveis e que a exposição é, em certo sentido, uniforme. Desta forma, R_0 é um número idealizado dado que nunca é o caso de que todas as pessoas dentro da população são igualmente suscetíveis. Exemplificando-se para o caso da COVID19, com o valor razoável de $R_0 = 2{,}5$ a imunidade coletiva seria de 60%. Porém, o cálculo dos limiares a partir dessa fórmula têm sido questionados na literatura (especificamente para o SARS-CoV-2 tem-se o estudo de Aguas et al., 2020). O principal argumento é que a fórmula só pode ser aplicada em experimentos com imunização aleatória e que, na prática, tem-se que levar em conta o chamado coeficiente de variação individual (CV) dado que o vírus tenderia a infectar os mais propensos a adoecer em um primeiro momento.

O papel do CV é contrário ao do número de reprodução básico (R_0): quanto maior o CV menor o limiar da imunidade coletiva e quanto maior o valor de R_0 maior o limiar. A existência do CV indica que os vetores mais fortes de propagação da doença são selecionados (imunizados) logo de início e o cenário de seleção dos vetores continuaria até se atingir um limiar muito mais baixo do que o obtido pela fórmula inicialmente proposta: "*In an epidemic that takes its natural course, by contrast, the virus very specifically infects the people that are most susceptible first. This removes all of the strongest vectors early on, and continues to selectively remove the vectors until the herd immunity threshold is reached*". Diante disso, pode-se concluir que em modelos projetados para infecções naturais, a variabilidade na suscetibilidade é um fator determinante do limiar de imunidade coletiva. Se houver variabilidade zero e todos forem igualmente suscetíveis, o número de reprodução efetiva é igual ao número de reprodução básico e o limite da imunidade coletiva é o mesmo que seria para vacinações distribuídas

aleatoriamente. Contanto que a variabilidade seja diferente de zero, o limite de imunidade de rebanho é menor do que seria para vacinações distribuídas aleatoriamente. Se a variabilidade for grande, o limite é muito mais baixo do que seria para vacinações distribuídas aleatoriamente e uniformemente.

Se o cenário atual da COVID19 não permitir a aplicação da fórmula usual do limiar, tem-se proposto o limiar para a imunidade de rebanho a partir da seguinte fórmula: $1 - (1/R_0)^{1/(1+CV^2)}$. Grosso modo, a literatura tem reportado valores para R_0 situados em torno de 2,5 para o SARS-CoV-2. Para o CV, têm-se as estimativas relativas ao SARS-CoV-1 para Singapura e Beijing, ambas em torno de 2,6. Pode-se ainda listar os números do CV para malária no Amazonas, de 1,8 e para tuberculose no Brasil de 3,3. Usando, apenas como ilustração, os valores de $R_0 = 3$ e CV = 2,6, o limiar da imunidade coletiva seria de 13%. Para o mesmo valor de $R_0 = 3$ e CV próximo ao da tuberculose no Brasil, o limiar seria de 9%. O artigo de Aguas et al (2020) calcula o coeficiente de variação (CV) da suscetibilidade para países europeus concluindo que o limiar da Europa se situa entre 10% e 20%. Em particular, "*The estimated thresholds for the countries they examined were 9.6-11% in Belgium, 20-21% in England, 6-7.3% in Portugal, and 11-12% in Spain.*"

Combinando as diferentes estimativas do número de infectados relatadas acima e os limiares obtidos pela fórmula padrão e pela análise feita por Aguas et al. (2020), temos vários cenários para avaliar a situação atual da epidemia no Brasil. Se admitirmos que a imunidade coletiva só é atingida em limiares altos (em geral acima de 50%) obtidos com a fórmula padrão, o Brasil ainda estaria longe de tal imunidade. Se o limiar estiver mais próximo dos valores propostos por Aguas et al. (2020), digamos em 20%, então a conclusão seria de que boa parte do Brasil já estaria próxima da imunidade coletiva. Exceto pelas estimativas mais baixas de taxa de infecção (de 3,7% e 8,4%, ambas obtidas com foco na subpopulação de casos sintomáticos), as demais estimativas para o número de infectados estão entre 10% e 20%. Cabe ainda observar que o número de infectados pode não ser o único valor a se considerar. Muito mais difícil de se mensurar (ainda mais para um vírus menos conhecido como o SARS-CoV-2) é a **imunidade pré-existente** e quantos indivíduos poderiam ser somados aos infectados simplesmente por terem imunidade "natural" ou cruzada (Doshi, 2020).

Por fim, dois estudos recentes, um da Universidade de São Paulo (Buss et al, *preprint*) e outro da Universidade Federal de Pelotas[4], apontam para a desaceleração da epidemia no Brasil com base na análise da presença de anticorpos na população[5] e,

[4] http://www.epidemio-ufpel.org.br/site/content/sala_imprensa/4-fase-do-epicovid19-mostra-desaceleracao-do-coronavirus-no-brasil.php?noticia=3149.

[5] "Diminuiu a proporção da população que apresenta anticorpos, o que confirma a desaceleração da epidemia na maior parte do país. Ao contrário do que se pensava no início da pandemia, os anticorpos detectáveis pelo teste duram apenas algumas semanas. Isso vem acontecendo em

possivelmente, corroboram os cenários aventados acima que combinam taxas de infecção altas e limiares de imunidade coletiva mais baixos.

4.2. PICOS/ONDAS DAS INFECÇÕES

O conceito de pico (e quantas ondas ocorrem em um dado período de uma epidemia) depende fortemente do número de infecções e, portanto, não depende unicamente dos casos confirmados, mas também do padrão de subnotificação que pode variar ao longo do tempo. Em geral, espera-se que a subnotificação diminua com a evolução da epidemia. Conforme reforçado por nossas estimativas, o número de casos confirmados não pode, por si só, ser usado para identificar a existência de uma segunda onda. Seria preciso estimar o número de infecções na época de referência (primeira onda), em algum(s) ponto(s) no interstício e, por fim, na presumida segunda onda. Isto deve ser feito para que não se obtenha uma falsa impressão da evolução da epidemia na região de interesse. A Figura 6 ilustra este ponto.

O gráfico (a) traz números (hipotéticos) de novos casos confirmados diários. Olhando somente para ele fica uma impressão clara de uma segunda onda, menos expressiva do que a primeira, mas, ainda assim, "relevante". Se a subnotificação for constante, digamos 10 vezes, ao longo de todo o período então a conclusão é a mesma e somente a escala muda.

O gráfico (b) assume que, por força de padrões de testagem, entre outros fatores mencionados aqui, a subnotificação esteja diminuindo ao longo do tempo: especificamente, o fator de subnotificação era de 10 vezes no início e foi decaindo a cerca de 2 vezes após o 100º dia. Neste caso não se tem mais a noção de uma segunda onda, mas sim de um leve aumento no número de casos, possivelmente decorrente de uma expansão da epidemia em sub-população previamente ilesa.

O gráfico (c) supõe o oposto (a subnotificação aumentou ao longo do tempo) e, portanto, a situação está piorando sobremaneira.

---

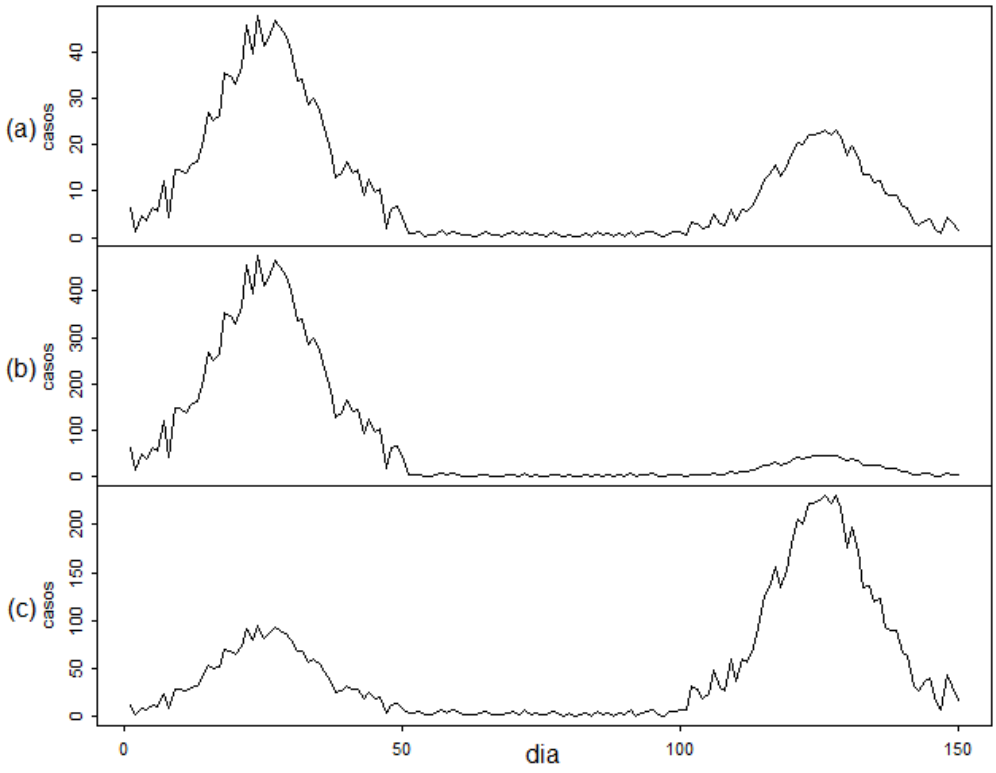
diversos países, com distintos tipos de testes de anticorpos (...). A queda em níveis de anticorpos ao longo do tempo não indica que os indivíduos deixem de estar protegidos, pois seus organismos guardam a memória imunológica para produzir anticorpos rapidamente em caso de uma nova infecção. Os indivíduos com testes positivos na última fase do EPICOVID-19 representam aqueles com infecções relativamente recentes. Muitas pessoas que foram infectadas há mais tempo passaram a apresentar resultados negativos atualmente. Portanto, não está correto usar a estimativa atual para indicar uma possível "imunidade de rebanho", tampouco para avaliar a probabilidade de uma "segunda onda" da pandemia." Informações para a imprensa - 4ª fase do EPICOVID19 - http://www.epidemio-ufpel.org.br/

Outro dado importante que pode ser complementar à análise de picos e ondas é o padrão de mortes em decorrência da COVID19. A não ser que tenha havido grandes avanços/retrocessos no tratamento ou que se possa mostrar que indivíduos do grupo de risco estejam se infectando a taxas menores/maiores[6], a existência de uma segunda onda deveria ser acompanhada de padrão similar nos óbitos. Caso haja variações como as citadas, então os padrões serão diferentes. Por exemplo, no momento da escrita deste artigo, boa parte da Europa apresenta padrão de casos confirmados similar ao gráfico (a) da Figura 6. Seria uma segunda onda? Se a subnotificação na época do primeiro pico era alta e agora está muito menor, a resposta seria negativa pois estaríamos em uma situação semelhante a Figura 6 (b). De fato, o número de mortes na suposta segunda onda está, até então, se comportando como em (b). Por outro lado, se a subnotificação aumentar (algo não esperado), então a distribuição dos casos deve apresentar o padrão (c).

Por fim, é importante que o conceito de segunda onda seja aplicado a uma mesma região e de tamanho compatível com o período analisado. Dados muito agregados, como por exemplo na Figura A1 são de pouca utilidade prática, uma vez que a suposta "segunda onda" pode representar a chegada da doença em uma região previamente não contaminada.

[6] Sem mencionar um terceiro fator, qual seja, mutação do vírus para uma forma mais/menos "agressiva".

Figura 6 – Padrões hipotéticos de evolução do número de infecções.



Fonte – Dados hipotéticos.

5. CONSIDERAÇÕES FINAIS

Estimar o número real de infectados é imprescindível para o entendimento da atual pandemia originada na China em menos de um ano, referente ao vírus SARS-CoV-2. No entanto, outros elementos precisam ser entendidos e associados a qualquer estimativa de número de infectados. É preciso entender/estimar o tamanho da imunidade pré-existente, estabelecer um limiar para a imunidade coletiva e acompanhar a evolução da subnotificação (que pode não ser a mesma durante o período analisado). Nossas estimativas de subnotificação se referem a agosto/2020 e estão nos níveis estadual e nacional. Para certos processos decisórios, talvez seja necessário informações municipais, se possível. As nossas estimativas de número de infectados são: **(i) estáticos** – precisam ser atualizados e monitorados ao longo do tempo para que sejam validados e interpretados corretamente; **(ii) agregados espacialmente** – apesar de trazerem informações não apenas para o Brasil mas em nível estadual, os dados precisam ser interpretados corretamente diante do tamanho e heterogeneidade de cada estado.

É extremamente importante que se continue o monitoramento dos infectados seja por meio de pesquisas como a PNAD-COVID19 ou por inquéritos sorológicos para que se consolidarem as estimativas das taxas de infecção a serem comparadas com os diferentes limiares de imunidade coletiva sendo publicados na literatura epidemiológica.

6. REFERÊNCIAS BIBLIOGRÁFICAS

AGUAS, Ricardo et al. *Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics of COVID-19 SARS-CoV-2*. 2020. Disponível como preprint no medRxiv e bioRxiv.

ANGRIST, J.; PISCHKE, J.-S. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, 2009.

BRASIL. Instituto Brasileiro de Geografia e Estatística. *PNAD COVID19: Informativo para a mídia*, 2020.

BUSS, Lewis F. et al. COVID-19 herd immunity in the Brazilian Amazon. Disponível como preprint em medRxiv 2020.09.16.20194787; doi: 10.1101/2020.09.16.20194787.

DOSHI, Peter. *Covid-19:* Do many people have pre-existing immunity? *BMJ*, 2020. 370:m3563.

GUDBJARTSSON, D. F. et a. Humoral Immune Response to SARS-CoV-2 in Iceland. *New England Journal of Medicine*. 2020 Sep 1. DOI: 10.1056/NEJMoa2026116. PMID: 32871063.

LAN, L. et al. Positive RT-PCR test results in patients recovered from COVID- 19. *Journal of the American Medical Association*. v. 323, p. 1502–1503, 2020.

LASH, T. L. et al. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer Science & Business Media, 2011.

KIYUKA, Patience K.; et al. Human Coronavirus NL63 Molecular Epidemiology and Evolutionary Patterns in Rural Coastal Kenya. *The Journal of Infectious Diseases*. v. 217, n. 11, p. 1728–1739, 2018. DOI: 10.1093/infdis/jiy098.

PEARCE, N. et al. Accurate statistics on COVID-19 are essential for policy guidance and decisions. *American Journal of Public Health*. v. 110, p. 949–951, 2020.

POOLE, David; RAFTERY, Adrian E. Inference for Deterministic Simulation Models: The Bayesian Melding Approach. *Journal of the American Statistical Association*. v. 95, n. 452, p. 1244-1255, 2000.

RAHMANDAD, Hazhir; LIM, Tse Yang; STERMAN, John. *Estimating COVID-19 Under-Reporting Across 86 Nations: Implications for Projections and Control*, 3 agosto 2020. Disponível na Internet em https://ssrn.com/abstract=3635047

RIBEIRO, Leonardo C.; BERNARDES, Américo T. Nota Técnica: Atualização da Estimativa de Subnotificação em Casos de Hospitalização por Síndrome Respiratória Aguda e Confirmados por Infecção por COVID-19 no Brasil e Estimativa para Minas Gerais. CEDEPLAR – UFMG, 2020.

YANG, Y. et al. *Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections*. 2020.

WU, Sean L. et al. Substantial underestimation of SARS-CoV-2 infection in the United States, *Nature Communications*. v. 11, n. 4507, 2020.

ANEXO 1 – ANÁLISE PROBABILÍSTICA SEMI-BAYESIANA DA CORREÇÃO DE VIÉS

A metodologia de correção de viés (Lash et al., 2011) é aplicável em estudos epidemiológicos que pretendem obter estimativas mais acuradas diante de testagens incompletas ou com elevada chance de reportar resultados falso-negativos. Pode-se corrigir uma ou múltiplas fontes de vieses. Em situações de múltiplas correções com maior complexidade faz-se necessário o uso de simulação, em geral, através de abordagem semi-Bayesiana e o estabelecimento de modelos probabilísticos (*a priori*) para as quantidades de interesse. A abordagem semi-Baeysiana de Wu et al. (2020) foi utilizada aqui com pequenas alterações na especificação das densidades *a priori*. Mantivemos as formas funcionais desenvolvidas pelos autores, mas foi necessário ajustar os parâmetros para valores condizentes com o observado no Brasil (ver Tabela A1 do Anexo 1). O método se mostrou robusto a variações razoáveis das especificações que fizemos.

Os principais elementos da modelagem *a priori* são: a probabilidade de ter sintomas moderados a severos entre os indivíduos testados, P(S_1|testado); probabilidade análoga para o grupo de indivíduos não testados, P(S_1|não testado); probabilidade de ter sintomas leves entre os testados com resultado positivo P(S_0|testado +), sensibilidade e especificidade dos testes para o COVID19.

Tabela A1: Especificações das densidades a priori, todas da família Beta truncada.

Distribuição	Min	Média	Max	DP
P(S_1|testado)	0,0000	0,5000	1,0000	0,2887
P(S_1|não testado)	0,0000	0,1500	0,3000	0,2000
α	0,7000	0,9000	1,0000	0,2000
β	0,0020	0,2000	0,5000	0,4000
P(S_0|testado +)	0,2500	0,7000	0,9000	0,4000
Sensibilidade	0,6500	0,8500	1,0000	0,3000
Especificidade	0,9800	0,9995	1,0000	0,0100

Fonte: elaboração própria.

Há ainda duas probabilidades (α e β) usadas na identidade probabilística

$$P(S_0|T^+) = \beta\big(1 - P(S_1|T^c)\big)/\Big(\beta\big(1 - P(S_1|T^c)\big) + \alpha P(S_1|T^c)\Big)$$

Na expressão acima, T é o evento "a pessoa foi testada para COVID19", T^c é o evento "a pessoa não foi testada para COVID19", T^+ é o evento "a pessoa testou positivo para COVID19", S_0 é o evento "a pessoa teve sintomas leves ou foi assintomática" e S_1

é o evento “a pessoa teve sintomas moderados ou severos”. Os parâmetros α e β.são, respectivamente, as razões de $P(T^+|S_1, T^c)$ e $P(T^+|S_0, T^c)$ com $P(T^+|T)$. Por termos especificações a priori em ambos os lados da expressão acima, a amostra conjunta do vetor $\big(P(S_0|T^+), \alpha, \beta, P(S_1|T^c)\big)$ é obtida por meio de “fundição” bayesiana (*Bayesian melding* – Poole e Raftery, 1999) através de 10^5 iterações do algoritmo de amostragem por importância e reamostragem (SIR). A correção por testagem incompleta é obtida a partir do número de infectados dentre os não testados com sintomas graves,

$$N^+_{T^c,S_1} = P(S_1|T^c) \times P(T^+|T^c) \times N_{T^c},$$

e de

$$N^+_{T^c,S_0} = \big(1 - P(S_1|T^c)\big) \times P(T^+|S_0, T^c) \times N_{T^c}.$$

N_{T^c} é a quantidade de pessoas não testadas na população. A correção do número de infectados pela acurácia dos testes é, portanto,

$$N^* = \frac{N^+ - \big(1 - S_p\big) \times N}{S_e + S_p - 1},$$

sendo N o tamanho da população, $N^+ = N^+_T + N^+_{T^c}$, N^+_T o número de casos confirmados dentre os testados, N^+_{Tc} a estimativa de casos positivos dentre os não testados , S_e a sensibilidade do teste (probabilidade do teste detectar a doença em uma pessoa infectada) e S_p a especificidade do teste (probabilidade do teste não detectar a doença em uma pessoa não infectada).

A proporção de infecções devido à não testagem pode ser separada em dois fatores: (1) a proporção atribuída às imperfeições nos testes

$$p_1 = \frac{N^* - N^+}{N^* - N^+_T}$$

e a proporção atribuída à testagem incompleta $p_2 = 1 - p_1$.

ANEXO 2: RESULTADOS ADICIONAIS

Criação de uma sub-amostra

Em muitos estados (municípios), o processo de retomada das atividades exigiu como pré-requisito a testagem dos trabalhadores. Essa exigência foi adotada na indústria, serviços e comércio. Normalmente, foi requisitada uma testagem "rápida" para essa categoria de trabalhadores. Diante disso, optou-se por selecionar o seguinte grupo de pessoas na amostra PNAD-COVID: a) indivíduos no mercado de trabalho, isto é, com idade entre 18 e 60 anos e; b) sem sintomas prévios. Com isso, espera-se que o viés de seleção relativo à procura do exame médico por se sentir doente seja eliminada. Nesse sentido, a proporção dos indivíduos infectados nessa sub-amostra poderia servir como um limite inferior para o número de infectados na população, dado que ele só poderia ser acrescido do número de resultados falso-negativos e dos sintomáticos acometidos pelo COVID19. Os resultados desse experimento indicam que o número mínimo (piso) de infectados no Brasil é de 14,3%, comparável ao obtido pela estimação semi-Bayesiana.

Tabela A2: Proporção de indivíduos assintomáticos que realizaram exame e obtiveram resultado positivo.

Região	Positivo	Negativo	Total	% Positivo	% Negativo
Centro-oeste	30.368	344.896	375.264	8,09%	91,91%
Nordeste	174.961	882.159	1.057.120	16,55%	83,45%
Norte	130.629	255.822	386.452	33,80%	66,20%
Sudeste	202.240	1.592.335	1.794.575	11,27%	88,73%
Sul	48.946	458.123	507.069	9,65%	90,35%
	587.145	3.533.335	4.120.481	14,25%	85,75%

Fonte: PNAD-COVID19, agosto de 2020.

Demais tabelas e figuras

Tabela A3: Proporção da população total infectada pelo COVID19 por estado – casos confirmados e estimativa após correção de viés.

UF	Observados	Estimados	% observado	% estimado
RR	40.385	169.656	7,28	30,56
AP	50.943	216.584	5,96	25,32
AC	22.669	172.826	2,58	19,65
AM	130.178	713.735	3,21	17,62
PA	271.456	1.522.930	3,13	17,59
SE	78.722	403.386	3,39	17,38
MA	184.923	1.173.353	2,61	16,56

AL	97.080	539.986	2,90	16,14
RO	37.511	261.642	2,10	14,63
CE	222.048	1.339.061	2,42	14,57
TO	45.820	229.722	2,90	14,55
MT	86.806	467.300	2,50	13,44
GO	210.667	919.906	2,96	12,91
ES	89.097	511.599	2,19	12,58
DF	130.253	375.711	4,26	12,29
PB	92.928	489.837	2,31	12,19
RN	71.738	381.859	2,03	10,79
SC	119.590	770.473	1,65	10,63
PE	127.652	998.418	1,33	10,44
BA	269093	1.480.846	1,80	9,93
PI	87.374	316.911	2,66	9,66
SP	858.917	4.410.417	1,85	9,52
RJ	294.058	1.628.878	1,69	9,38
PR	131.166	908.721	1,14	7,89
MG	213.968	1.601.270	1,00	7,52
MS	32.845	202.390	1,20	7,38
RS	107.654	663.780	0,94	5,82
Total	4.105.541	22.871.197	1,94	10,82

Fonte: Dados da PNAD-COVID19, agosto de 2020.

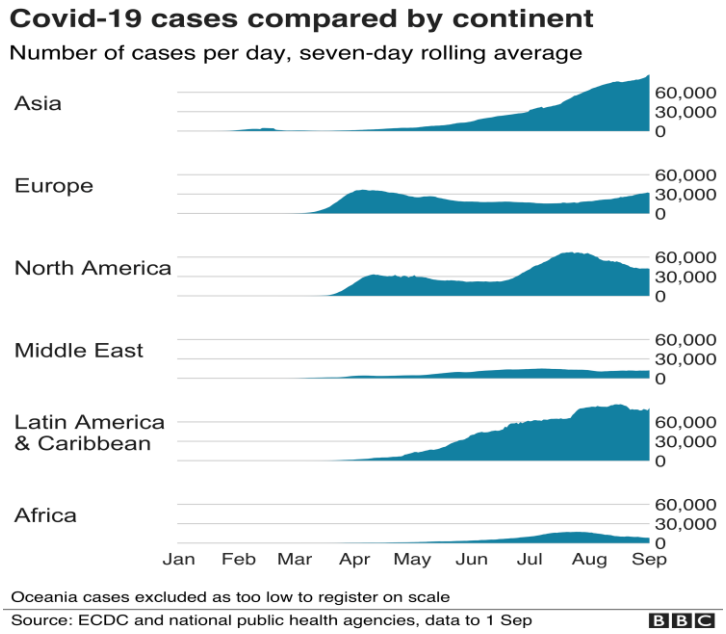
Tabela A4: Estimativa de subnotificação de casos confirmados de COVID19 por estado

UF	Observados	Estimados	Fator de correção	Intervalo FC
PE	127.652	998.418	7,82	4,06 – 12,16
AC	22.669	172.826	7,62	3,98 – 11,83
MG	213.968	1.601.270	7,48	3,90 – 11,62
RO	37.511	261.642	6,98	3,69 – 10,80
PR	131.166	908.721	6,93	3,66 – 10,73
SC	119.590	770.473	6,44	3,45 – 9,95
MA	184.923	1.173.353	6,35	3,43 – 9,79
RS	107.654	663.780	6,17	3,32 – 9,51
MS	32.845	202.390	6,16	3,33 – 9,51
CE	222.048	1.339.061	6,03	3,30 – 9,28
ES	89.097	511.599	5,74	3,17 – 8,82
PA	271.456	1.522.930	5,61	3,11 – 8,61
AL	97080	539.986	5,56	3,09 – 8,53
RJ	294.058	1.628.878	5,54	3,07 – 8,50
BA	269.093	1.480.846	5,50	3,05 – 8,44
AM	130.178	713.735	5,48	3,05 – 8,40
MT	86.806	467.300	5,38	3,00 – 8,24
RN	71.738	381.859	5,32	2,97 – 8,14
PB	92.928	489.837	5,27	2,94 – 8,05
SP	858.917	4.410.417	5,13	2,88 – 7,83
SE	78.722	403.386	5,12	2,89 – 7,81
TO	45.880	229.722	5,01	2,83 – 7,63
GO	210.667	919.906	4,37	2,54 – 6,54
AP	50.943	216.584	4,25	2,49 – 6,35

RR	40.385	169.656	4,20	2,47 – 6,27
PI	87.374	316.911	3,63	2,20 – 5,34
DF	130.253	375.711	2,88	1,87 – 4,15
Total	4.105.541	22.871.197	5,57	3,08 – 8,53

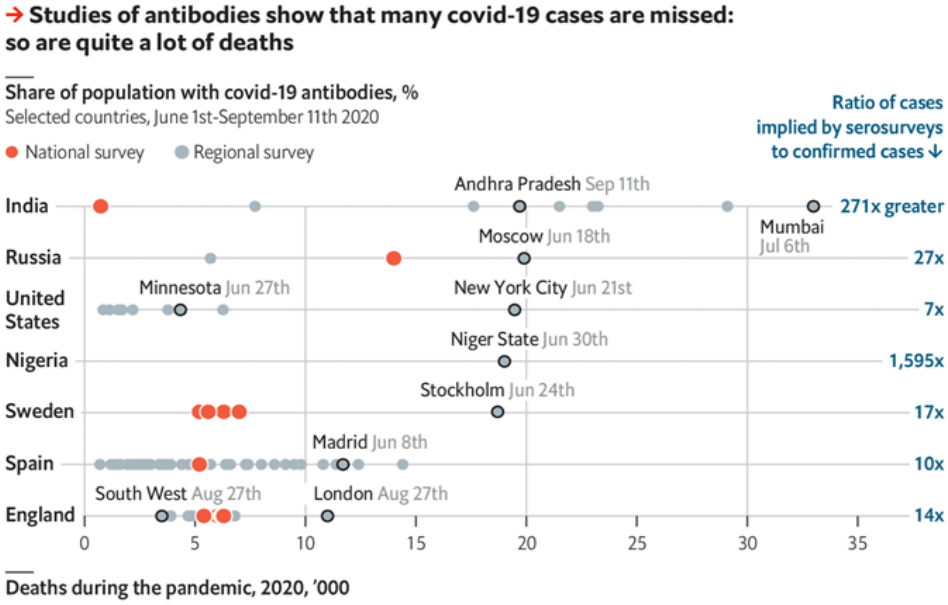
Fonte: Dados da PNAD-COVID19, agosto de 2020.

Figura A1: Números de casos por dia nos continentes.



Fonte: BBC.

Figura A2: Subestimação de casos de COVID19.



Fonte: The Economist, Sep24, 2020.