Supplemental Material: Modelling the effect of conditionally repeating hemoglobin measurements prior to blood donation

Mart Pothast, Katja van den Hurk & Mart Janssen

1 Introduction

This supplemental document contains details on deriving the effect of conditionally repeating hemoglobin measurements.

Blood donors have their hemoglobin (Hb) measured prior to every donation. Donation may only take place if Hb is greater than or equal to 7.8 mmol/L for females or 8.4 mmol/L for males. A drop of blood is collected via a prick in the finger and measured with a HemoCue apparatus. If the first measurement is below the threshold for donation the measurement is repeated and again a third time if the second measurement is still below.

To assess the variance of the measured Hb it is necessary to understand the distribution of Hb measurements that arises because of these conditionally repeated measurements. Given some assumptions about the underlying (true) population Hb distribution it is possible to derive the probability density function (pdf) of the second (and third) conditional measurement. Also, the pdf of the combination of measurements (as is eventually stored in the data) is derived.[1]

2 Assumptions

We will assume the underlying population hemoglobin pdf is a normal distribution with mean μ_{pop} and standard deviation σ_{pop}:

$$\text{Hb} \sim \mathcal{N}(\mu_{\text{pop}}, \sigma_{\text{pop}}), \tag{1}$$

where $\mathcal{N}$ exemplifies the normal distribution. μ_{pop} is different for males and females, but for σ_{pop} this is unclear: they are probably similar.

Furthermore, assume the measurement of Hb adds Gaussian noise so that:

$$\text{Hb}_{\text{meas}} \sim \mathcal{N}(\text{Hb}_{\text{true}}, \sigma_{\text{meas}}) \tag{2}$$

[1] All code can be found at `https://github.com/martp91/repeated_Hb_paper`.

and σ_{meas} is the standard deviation of the measurement of Hb. The distribution of the first measurement of Hb has a standard deviation of

$$\sigma_{\text{tot}} = \sqrt{\sigma_{\text{pop}}^2 + \sigma_{\text{meas}}^2}, \tag{3}$$

so that the pdf of the first measurements is

$$f_1(x_1) = \mathcal{N}(x_1; \mu_{\text{pop}}, \sigma_{\text{tot}}) \tag{4}$$

3 Second measurement

A second measurement for any first measured would still be a distribution as in Equation (2). Because these measurements are from the same 'true Hb', they are correlated and their covariance is

$$\text{Cov}[x_1, x_2] = \sigma_{\text{pop}}^2, \tag{5}$$

and their correlation coefficient

$$\rho = \frac{\sigma_{\text{pop}}^2}{\sigma_{\text{tot}}^2}. \tag{6}$$

Another way to appreciate the two Hb measurements is as a bivariate normal distribution:

$$g_{2\text{D}}(x_1, x_2) = \frac{1}{2\pi\sigma_{\text{tot}}^2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(x_1'^2 + x_1'^2 - 2\rho x_1' x_2'\right)\right], \tag{7}$$

with $x_i' = (x_i - \mu_{\text{pop}})/\sigma_{\text{tot}}$.

3.1 Conditional distribution

Since the second measurement is only done if the first measurement is below the cutoff c, the probability of the second measurement is

$$P(x_2|x_1 < c) = \frac{P(x_2 \cap x_1 < c)}{P(x_1 < c)}. \tag{8}$$

The probability is the partial integral over Equation (7) normalised by the cdf of the (univariate) normal distribution

$$P(x_2|x_1 < c) = \int_{-\infty}^{c} f_{2\text{D}}(x_1, x_2)dx_1 \Big/ \int_{-\infty}^{c} \mathcal{N}(x_1)dx_1 \tag{9}$$

This integral has an analytical solution using the error function such that the pdf of the conditional second measurement is

$$f_2(x_2') \equiv P(x_2|x_1 < c) = \frac{\left[1 + \mathrm{erf}\left(\frac{c' - x_2'\rho}{\sqrt{2-2\rho^2}}\right)\right]}{\sqrt{2\pi}\sigma_{\mathrm{tot}}\left[1 + \mathrm{erf}\left(\frac{c'}{\sqrt{2}}\right)\right]} e^{-\frac{x_2'^2}{2}} \tag{10}$$

$$= \frac{\Phi\left(\frac{c' - \rho x_2'}{\sqrt{1-\rho^2}}\right)}{\Phi\left(c'\right)} \frac{\phi\left(x_2'\right)}{\sigma_{\mathrm{tot}}} \tag{11}$$

with $x_2' = (x_2 - \mu_{\mathrm{pop}})/\sigma_{\mathrm{tot}}$ and $c' = (c - \mu_{\mathrm{pop}})/\sigma_{\mathrm{tot}}$.

The expectation value of the second measurement is given by:

$$E[x_2] = \mu - \rho\sigma_{\mathrm{tot}} \frac{\phi(c')}{\Phi(c')} \tag{12}$$

3.2 Difference between first and second measurement

Given a sample of two subsequent measurements it is possible to determine the measurement uncertainty σ_{meas} using

$$\mathrm{Var}[\mathrm{x}_1 - \mathrm{x}_2] = \mathrm{Var}[\mathrm{x}_1] + \mathrm{Var}[\mathrm{x}_2] - 2\mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2]. \tag{13}$$

For a non-conditional second measurement this would simplify to

$$\mathrm{Var}[\mathrm{x}_1 - \mathrm{x}_2] = 2\sigma^2_{\mathrm{meas}}. \tag{14}$$

For the conditional second measurement the variance is underestimated. Using the pdf from Equation (8) and integrating[2] (or performing a simple Monte Carlo) to get the variance and covariance it is possible to calculate by how much the variance is underestimated.

Computing the correction on standard deviation

The standard deviation of the difference between the conditional measurements $\sigma_{\mathrm{diff,cut}}$ is calculated for given μ_{pop}, σ_{pop} and cutoffs. For males $\mu_{\mathrm{pop}} = 9.4$ and the cutoff $c = 8.4$, for females $\mu_{\mathrm{pop}} = 8.5$ and $c = 7.8$, all in mmol/L. $\sigma_{\mathrm{pop}} \in [0.4, 0.5, 0.6, 0.7]$ and $\sigma_{\mathrm{meas}} \in [0.1, 0.2...1.0]$ in steps of 0.1 are used. Results are shown in Figure 1.

For $\sigma_{\mathrm{diff,cond}} \leq 0.2$ there is no difference. Above this value the points fall on a straight line (at least up to 1.0). There is almost no difference for different μ_{pop} and c. The slopes of the fitted lines in Figure 1 depend slightly on σ_{pop}.

The following function is fitted to the points in Figure 1

$$\sigma_{\mathrm{meas}} = \begin{cases} \sigma_{\mathrm{diff,cond}}, & \text{for } \sigma_{\mathrm{diff,cond}} \leq 0.2 \\ 0.2 + (\sigma_{\mathrm{diff,cond}} - 0.2)(1.46 - 0.4\sigma_{\mathrm{pop}}) & \text{otherwise.} \end{cases} \tag{15}$$

[2] The integrals are over the bivariate normal Equation (7) of the form: $\int_{-\infty}^{c}\int_{-\infty}^{\infty} g_{2D}(x_1, x_2)dx_1dx_2$.

This function can be used to estimate the true σ_{meas} from the observed standard deviation of the difference between conditionally repeated measurements. The difference due to μ_{pop} and c is less than 0.01 for reasonable values.

When σ_{pop} is unknown it can be estimated by choosing an appropriate guess and iterating a few times using Equation (15) and taking the sample standard deviation with Equation (3).

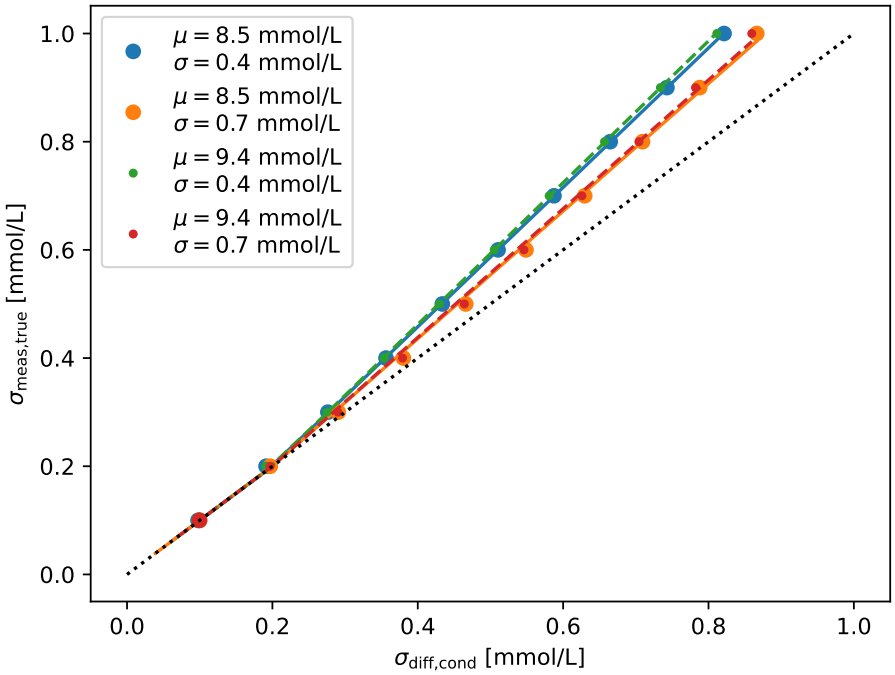

Figure 1: Computed standard deviation of the difference between first and second conditional measurement (on the x-axis) for given true measurement standard deviation (y-axis). Coloured lines represent straight lines fitted to the points. The dotted line is 1:1.

4 Third conditional measurement

The measurement is repeated a third time if the second measurement is also below the cutoff. The probability then is

$$P(x_3|(x_1 < c \cap x_2 < c)) = \frac{P(x_3 \cap (x_1 < c \cap x_2 < c))}{P(x_1 < c \cap x_2 < c)}. \tag{16}$$

Which corresponds to integrating over a 3D Gaussian ('trivariate normal') with the same covariance and variance as in the 2D case (i.e. the correlation between

the first, second and third measurement is the same). Giving the pdf of the conditional third measurement as

$$f_3(x_3) \equiv P(x_3|(x_1 < c \cap x_2 < c)) = \int_{-\infty}^{c}\int_{-\infty}^{c} f_{3\mathrm{D}}(x_1, x_2, x_3)dx_1dx_2 \Big/ \int_{-\infty}^{c}\int_{-\infty}^{c} f_{2\mathrm{D}}(x_1, x_2)dx_1dx_2 \ , \tag{17}$$

which can be integrated numerically.

The calculated distributions of the second and third measurements have been verified with 10^6 Monte Carlo simulated measurements (see Figure 2). See also Figure 3 for the 2D distribution of the first and second measurement.

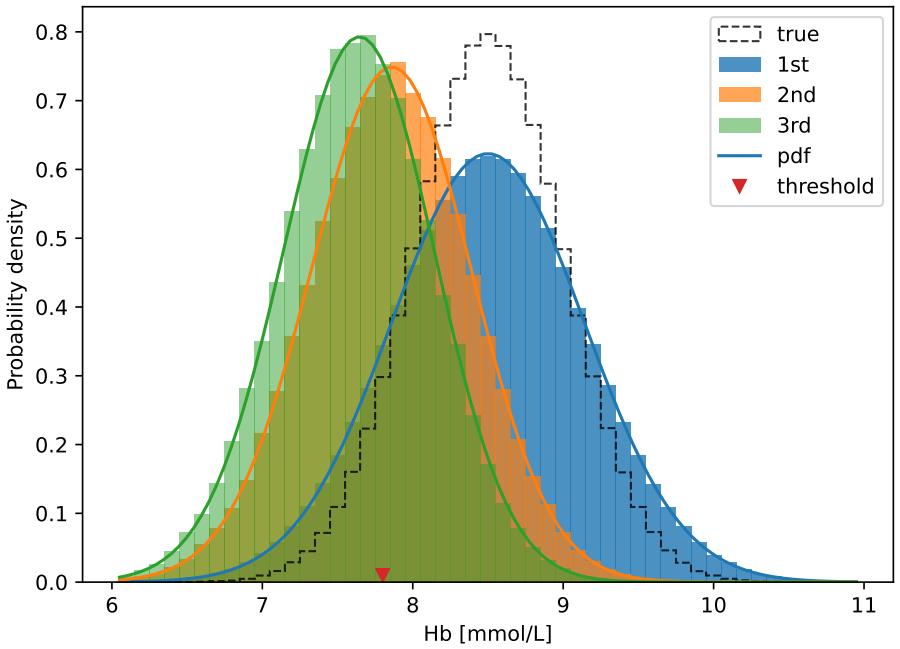

Figure 2: Simulated first, second and third conditional measurements (histogram as stepped lines) and the calculated pdfs as a smooth line. The following values are used: $\mu_{\mathrm{pop}} = 8.5$, $\sigma_{\mathrm{pop}} = 0.5$, $\sigma_{\mathrm{meas}} = 0.4$ and $c = 7.8$—all in mmol/L.

5 Combination of measurements

We will assume the algorithm for taking the conditional measurements is as in Algorithm 1. The combined pdf for any measurement from the algorithm is

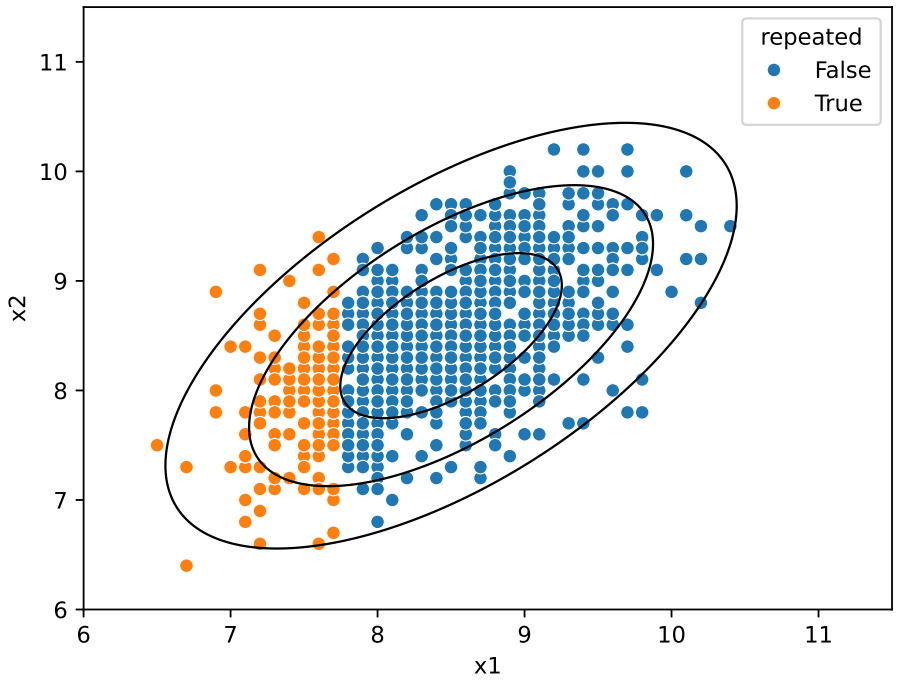

Figure 3: Scatter plot of 1500 once repeated measurements (x_1 the first and x_2 the second) with the parameters as in Figure 2. The points that would have been repeated had $x_1 < c$ are marked in orange. The contour lines show the 50%, 90% and 99% of the bivariate normal distribution.

then:

$$\begin{aligned}\text{pdf}(x) = (1-p_1)f_1^{\text{trunc}}(x) + p_1((1-p_2)f_2^{\text{trunc}}(x) + \\ p_1p_2(1-p_3)f_3^{\text{trunc}}(x) + p_1p_2p_3f_3^{\max}(x)\end{aligned} \tag{18}$$

Here f_i^{trunc} are the pdf's of the first, second and third measurement from Equations (4), (11) and (17) truncated from c to ∞. Likewise, p_i are the probability that the i-th measurement is below the cutoff c (acquired by integrating $f_i(x)$). $f_3^{\max}$ is the pdf of the maximum of the three measurements truncated from $-\infty$ to c, which is given by

$$f_3^{\max}(x) = \frac{d}{dx}\int_{-\infty}^{x}\int_{-\infty}^{x}\int_{-\infty}^{x} f_{3D}(x,x,x)dxdydz \tag{19}$$

where the integral represents the cdf of the trivariate normal distribution with the covariance as before and the derivative gives the pdf. Additional probability that a measurement is in fact repeated can be added as a multiplicative factor on the desired pdf component.

Algorithm 1 Algorithm for three conditional measurements

for i in 1, 2, 3 **do**
  $x_i \sim \mathcal{N}(x_{\text{true}}, \sigma_{\text{meas}})$
  if $x_i \geq c$ **then return** x_i
  end if
end for
return $\max x_i$

The calculated pdf from Equation (18) is verified by MC simulations and shown in Figure 4.

Note, that if the measurements are rounded, c (as the lower/upper limit of integration) has to be adjusted, e.g. when rounded to one decimal $c \to c - 0.05$.

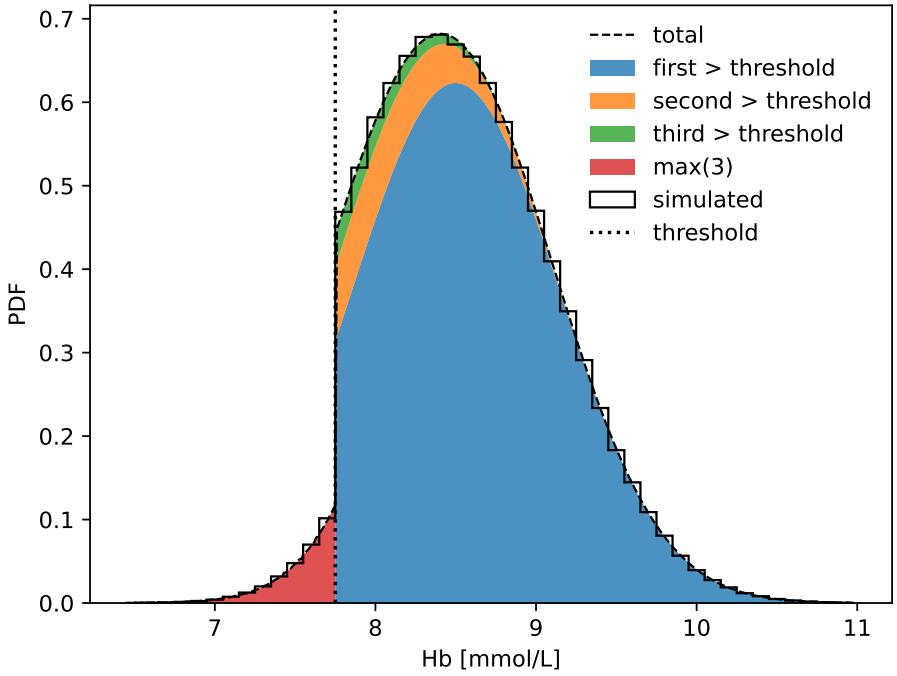

Figure 4: Combination of three measurements (see Figure 2) following the algorithm from Algorithm 1. The continuous line is the pdf as calculated with Equation (18) and the histogram is from 10^6 MC simulations. Same parameters as in Figure 2.

6 Likelihood

It is also possible to consider the likelihood that a single measurement x (being one of the three conditional fingerprick Hb measurements done) is from a true Hb, denoted θ. In this case, given the measurement algorithm as in Algorithm 1, this simplifies considerably:

$$\mathcal{L}(\theta|x) = \begin{cases} \mathcal{N}(x, \theta, \sigma_{\text{meas}})(1 + p + p^2) & \text{if } x \geq c \\ p^3 f_3^{\max}(x, \theta, \sigma_{\text{meas}}) & \text{if } x < c, \end{cases} \tag{20}$$

with $p = \int_{-\infty}^{c} dx \mathcal{N}(x, \theta, \sigma_{\text{meas}})$ the probability that a measurement is below the cutoff and

$$f_3^{\max}(x) = 3\mathcal{N}(x, \theta, \sigma_{\text{meas}}) F(x, \theta, \sigma_{\text{meas}})^2. \tag{21}$$