After my talk on Thursday, some people asked how to improve Hoeffding's inequality, and how it could be made to give confidence-intervals that do not exceed the data's bounds. Hoeffding's inequality is easy to use and can be essential in some formal proofs. But my research has led me to know that there are better bounds. And if my learning is usefull for someone else, then I regard that as a best kind of outcome!

The easiest way to show how Hoeffding's inequality can be improved is to show how a better bound can be created by omitting an approximation in its derivation - ie. creating a universally more powerfull bound.
Hoeffding's inequality is an example of a Chernoff bound, which uses Markov's inequality:

Lemma 1 (Markov's Inequality). *for any non-negative random variable X and any $a > 0$ that: $\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a$*

Lemma 2 (Chernoff Bound). *If $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is sample mean of n independent and identically distributed samples of random variable X ($x_i \sim X$), then for any $s, t > 0$: $\mathbb{P}(\hat{\mu} \geq t) \leq \mathbb{E}\left[\exp(sX)\right]^n \exp(-snt)$*

Proof. $\mathbb{P}(\hat{\mu} \geq t) = \mathbb{P}\left(\exp\left(s\sum_{i=1}^{n} x_i\right) \geq \exp(snt)\right)$ hence by Markov's inequality $\mathbb{P}(\hat{\mu} \geq t) \leq \mathbb{E}\left[\exp\left(s\sum_{i=1}^{n} x_i\right)\right]\exp(-snt)$
The result follows as we assume that our samples are independant (for any independant variables A, B that $\mathbb{E}[AB] = \mathbb{E}[A]\,\mathbb{E}[B]$). □

Theorem 1 (Hoeffding's inequality for mean zero). *Let X be a random variable that is bounded $a \leq X \leq b$, with a mean $\mu = 0$. Then letting $D = b - a$, then for any $t > 0$, the mean $\hat{\mu}$ of n independent samples of X is bounded: $\mathbb{P}(\hat{\mu} \geq t) \leq \exp\left(\frac{-2nt^2}{D^2}\right)$*

Proof. To prove Hoeffding's inequality we develop an upper bound for $\mathbb{E}[\exp(sX)]$, if we assume variable X has a probability density function $f(x)$, then we can fit a line over $\exp(sx)$ as: $\mathbb{E}[\exp(sX)] = \int_a^b f(x)\exp(sx)dx \leq \int_a^b f(x)(\frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa})dx$
Using the fact that the mean $\mu = \int_a^b f(x)xdx = 0$ thus: $\mathbb{E}[\exp(sX)] \leq \frac{1}{sb-sa}\left(sb\exp(sa) - sa\exp(sb)\right)$
Given the fact that for any κ, γ: $\frac{1}{\kappa-\gamma}(\kappa\exp(\gamma) - \gamma\exp(\kappa)) \leq \exp\left(\frac{1}{8}(\kappa-\gamma)^2\right)$ (1)
Thus $\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{1}{8}s^2(b-a)^2\right)$ and by our Chernoff bound (lemma 2) we get: $\mathbb{P}(\hat{\mu} \geq t) \leq \exp\left(\frac{1}{8}s^2(b-a)^2n - snt\right)$ And minimising with respect to s gives the result. □

At a first glance, the most limiting feature of this derivation is the requirement that the mean is zero, however this is ultimately immaterial and simplifies the derivation. We just consider our data as if it were shifted to have a mean of zero, leaving D unchanged. hence we get the equation we know and love:

Theorem 2 (Hoeffding's inequality). *Let X be a real-valued random variable that is bounded $a \leq X \leq b$. Then for $D = b - a$ and any $t > 0$, the mean $\hat{\mu}$ of n independent samples of X is probability bounded by:*

$$\mathbb{P}(\hat{\mu} - \mu \geq t) \leq \exp\left(-2nt^2/D^2\right) \quad \textit{Or by rearranging:} \quad \mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{D^2\log(1/t)/(2n)}\right) \leq t \tag{2}$$

However we can easily do better.

Theorem 3. *Let X be a real-valued random variable that is bounded $a \leq X \leq b$, with a mean μ of zero. Then for $t > 0$, the mean $\hat{\mu}$ of n independent samples of X is probability bounded by:*

$$\mathbb{P}(\hat{\mu} \geq t) \leq \left(\frac{b}{b-a}\left(\frac{b(a-t)}{a(b-t)}\right)^{\frac{a-t}{b-a}} - \frac{a}{b-a}\left(\frac{b(a-t)}{a(b-t)}\right)^{\frac{b-t}{b-a}}\right)^n \tag{3}$$

Proof. Exactly the same proof as Theorem 1 except do not use Equation 1, leading to:
$\mathbb{P}(\hat{\mu} \geq t) \leq (b\exp(sa) - a\exp(sb))^n \exp(-snt)(b-a)^{-n}$ and minimising with respect to s gives the result. □

Now, this concentration inequality is more powerful but more difficult to manipulate, we also dont usually know a or b (since that would correspond to knowing the mean itself). But we usually know that our data is bounded, for instance if we are dealing with binary data then we know that, our variable X (not shifted to have a zero mean) is bounded $0 < X < 1$, and we directly get:

Theorem 4 (Also called Hoeffding's inequality). *Let X be a real-valued random variable that is bounded $0 \leq X \leq 1$, with mean μ. Then for $t > 0$, the mean $\hat{\mu}$ of n independent samples of X is probability bounded by:*

$$\mathbb{P}(\hat{\mu} - \mu \geq t) \leq \left[\left(\frac{1-\mu}{1-t-\mu}\right)^{1-t-\mu}\left(\frac{\mu}{t+\mu}\right)^{t+\mu}\right]^n \tag{4}$$

Proof. Follows directly from Theorem 3 with the substitution $a = -\mu$ and $b = 1 - \mu$. □

This eqution is also credited to Hoeffding and is found widely in literature. So, if you have binary data and your sample mean $\hat{\mu}$ and you want know how likely it is at underestimating the data mean μ by t then this equation gives this (the equivalent inequality for overestimation is similar)

So, how much improvement is given by Theorem 4 over our Hoeffding's inequality Theorem 2 over what might be achieved perfectly?
Consider the following figure:

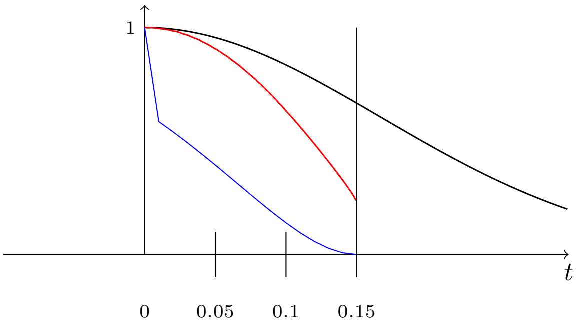

Figure 1: for binary data, if your sample mean is $\hat{\mu} = 0.85$, the probability envelopes of Hoeffdings inequality (black) and from Theorem 4 (red) above a hypothetical minimum (blue) for $n = 9$

And from this figure, we can see that the bound derived from Theorem 4 as the red line is below the black line corresponding to Hoeffding's inequality (Theorem 2). The ideal blue bound (or atleast what I currently suspect is ideal (!) - contact me if you might like to do collaboration or something) is as follows:

$$\mathbb{P}(\hat{\mu} - \mu > t) \leq \max\left(\left(1 + \frac{t}{\hat{\mu} - 1}\right)^n, \sum_{m=0}^{\lfloor \hat{\mu} n \rfloor} \binom{n}{m} \left(\frac{(\hat{\mu} + t - 1)(n - \lfloor \hat{\mu} n \rfloor)}{n(1 - \hat{\mu})}\right)^{n-m} \left(1 - \frac{(\hat{\mu} + t - 1)(n - \lfloor \hat{\mu} n \rfloor)}{n(1 - \hat{\mu})}\right)^m\right)$$

Which I anticipate is the result of Optimal Uncertainty Quantification procedure [1] - which basically is the process of searching directly for the worst case probability distribution for your random variables constrained by what you know.

In anycase, the literature on concentration inequalities is simply vast, and there are lots of improvements to be made, and an array of techniques which can be applied. Sticking to the original Hoeffding's inequality is safe for publication purposes - as people understand and know it. but there is no reason to stay bound to what we are familiar with.

1 an additional note

So, in deriving (most, but importantly not all) concentration inequalities (such as Hoeffding's) we start with the assumption of distribution parameters and then deduce the likely error of sampling statistics from that. However the way in which we often use these inequalities is in reverse, starting from knowledge of sampling statistics and then making inferences about the distribution parameters. And if you are a baysian like I am, you should be very worried... as this is technically invalid use. However, I understand that ultimately the same problem is part of statistical hypothesis testing generally.

References

[1] H. Owhadi, C. Scovel, T. Sullivan, M. McKerns, and M. Ortiz. Optimal uncertainty quantification. *SIAM Review*, 55(2):271–345, 2013.