Further Tail Bounds: Chernoff Bounds

Josep Díaz   Maria J. Serna   Conrado Martínez
U. Politècnica de Catalunya

RA-MIRI 2023–2024

Why do we need more concentration bounds?

- Remember that given a random variable, we are trying to determine how concentrated it is, trying to show the probability of hitting a random instance which deviates far from the expectation μ, is small.
- We aim to have random variables (events) which are concentrated around its mean with high probability.
- We saw that if $X \geqslant 0$ Markov can give an indication that there are values very far away from its mean, but in general is too weak for proving strong concentration results.
- Chebyshev's inequality can give stronger results for concentration of X around μ, but we must compute $\mathbb{V}[X]$, which could be difficult.

Chernoff Bounds

H. Chernoff (1923–)

Sergei Bernstein (1924), Wassily Hoeffding (1964), Herman Chernoff (1952)

The Chernoff bound can be used when the random variable X is the sum of several independent random variables. For Bernouilli trials, where each X_i can have probability of success p_i. The particular case where all p_i are equal is the Bernouilli trials, Chernoff bound takes the form given in the following Theorem.

Chernoff Bounds

Theorem (Theorem 1)

Let $\{X_i\}_{i=0}^{n}$ *be independent Bernouilli trials, with* $\mathbb{P}[X_i = 1] = p_i$*. Then, if* $X = \sum_{i=1}^{n} X_i$*, and* $\mu = \mathbb{E}[X]$*, we have*

1. $\mathbb{P}[X \leqslant (1-\delta)\mu] \leqslant \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$*, for* $\delta \in (0, 1)$*.*
2. $\mathbb{P}[X \geqslant (1+\delta)\mu] \leqslant \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$ *for any* $\delta > 0$*.*

Weak Chernoff bound

Corollary (Corollary 2)

Let $\{X_i\}_{i=0}^n$ *be independent Bernouilli trials, with* $\mathbb{P}[X_i = 1] = p_i$*. Then if* $X = \sum_{i=1}^n X_i$*, and* $\mu = \mathbb{E}[X]$*, we have*

1. $\mathbb{P}[X \leqslant (1-\delta)\mu] \leqslant e^{-\mu\delta^2/2}$*, for* $\delta \in (0,1)$*.*
2. $\mathbb{P}[X \geqslant (1+\delta)\mu] \leqslant e^{-\mu\delta^2/3}$*, for* $\delta \in (0,1]$*.*

Corollary (Corollary 3)

Let $\{X_i\}_{i=0}^n$ *be independent Bernouilli trials, with* $\mathbb{P}[X_i = 1] = p_i$*. Then if* $X = \sum_{i=1}^n X_i$*,* $\mu = \mathbb{E}[X]$ *and* $\delta \in (0,1)$*, we have*

$$\mathbb{P}[|X-\mu| \geqslant \delta\mu] \leqslant 2e^{-\mu\delta^2/3}.$$

Proof of Corollary 3

Sketch

Using Cor. 2

$$\begin{aligned}\mathbb{P}[|X-\mu| \geqslant \delta\mu] &= \mathbb{P}[X < (1-\delta)\mu] + \mathbb{P}[X \geqslant (1+\delta)\mu] \\ &\leqslant e^{-\mu\delta^2/2} + e^{-\mu\delta^2/3} \leqslant 2e^{-\mu\delta^2/3}\end{aligned}$$

$\square$

Proof of Corollary 2.1

Proof

For Cor. 2.1: Using Thm. 1.1, we must prove that, for $\delta \in (0,1)$, we have $\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \leqslant e^{-\mu\delta^2/2} = \left(e^{-\delta^2/2}\right)^{\mu}$.

Let $f(\delta) = \ln\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right) - \ln\left(e^{-\delta^2/2}\right)$
$f(\delta) = -\delta - (1-\delta)\ln(1-\delta) + \delta^2/2 \leqslant 0$.

Differenciating $f(\delta)$:

$$f'(\delta) = \ln(1-\delta) + \delta$$
$$f''(\delta) = \frac{-1}{1-\delta} + 1 \leqslant 0$$

$\Rightarrow f''(\delta) < 0$ in $(0,1)$ and as $f'(0) = 0$, then $f'(\delta) \leqslant 0$ in $[0,1)$, i.e. $f(\delta)$ is non-increasing in $[0,1)$.

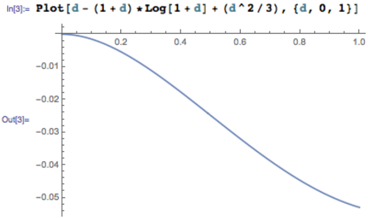
As $f(0) = 0 \Rightarrow f(\delta) \leqslant 0$ for $\delta \in (0,1)$. $\square$

Proof of Corollary 2.2

Proof

For Cor. 2.2: Using Thm. 1.2, we must prove that for $\delta \in (0,1)$, we have $\left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \leqslant e^{-\delta^2/3}$.

Taking logs: $f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \delta^2/3 \leqslant 0$. Differentiating 2 times $f(\delta)$, and using the same argument as above, we see $f(\delta) \leqslant 0$ in $(0,1]$. $\square$



An easy application

Back to an old example: We flip n-times a fair coin, we wish an upper bound on the probability of having at least $\frac{3n}{4}$ heads.
Recall Let $X \sim \text{Bin}(n, 1/2)$, then, $\mu = n/2$, $\mathbb{V}[X] = n/4$.

We want to bound $\mathbb{P}\left[X \geqslant \frac{3n}{4}\right]$.

- Markov: $\mathbb{P}\left[X \geqslant \frac{3n}{4}\right] \leqslant \frac{\mu}{3n/4} = 2/3$.
- Chebyshev: $\mathbb{P}\left[X \geqslant \frac{3n}{4}\right] \leqslant \mathbb{P}\left[|X - \frac{n}{2}| \geqslant \frac{n}{4}\right] \leqslant \frac{\mathbb{V}[X]}{(n/4)^2} = \frac{4}{n}$.
- Chernoff: Using Cor. 2.2,
  $\mathbb{P}\left[X \geqslant \frac{3n}{4}\right] = \mathbb{P}\left[X \geqslant (1+\delta)\frac{n}{2}\right] \Rightarrow (1+\delta)\frac{3}{2} \Rightarrow \delta = \frac{1}{2}$
  $\implies \mathbb{P}\left[X \geqslant \frac{3n}{4}\right] \leqslant e^{-\mu\delta^2/3} = e^{-\frac{n}{24}}$

Example

- If $n = 100$, Cheb. $= 0.04$, Chernoff $= 0.0155$
- If $n = 10^6$, Cheb. $= 4 \times 10^{-6}$, Chernoff $= 2.492 \times 10^{-18095}$

Another example

Example

Toss n times a fair coin, what is the probability of deviating from $n/2$ heads?

Let $X = \#$ heads, then $\mu = n/2$ and $\mathbb{V}[X] = n/4$.

1. Markov: $\mathbb{P}[X \geqslant n/2] \leqslant \frac{n/2}{n/2} = 1$. So $\mathbb{P}[X \leqslant n/2] \geqslant 0$. $\implies$ no information!
2. Chebyshev: Between $n/4$ and $3n/4$ heads:
$\mathbb{P}\left[|X - \frac{n}{2}| \geqslant \frac{n}{4}\right] \leqslant \frac{4}{n}$
3. Chernoff: Using the last bound
$\mathbb{P}\left[|X - \frac{n}{2}| \geqslant \frac{1}{2}\sqrt{6n \ln n}\right] \leqslant 2e^{-\frac{1}{3}\frac{n}{2}\frac{6 \ln n}{n}} = \frac{2}{n}$
Even $\mathbb{P}\left[|X - \frac{n}{2}| \geqslant \frac{n}{4}\right] \leqslant 2e^{-\frac{1}{3}\frac{n}{2}\frac{1}{4}} \leqslant 2e^{-\frac{n}{24}}$

Proof of Theorem 1: Upper tail

Proof

Note if for a r.v. X, and $a > 0$ and for any $t > 0$ we have

$$\boxed{(e^{tX} \geqslant e^{ta}) \Leftrightarrow (X \geqslant a)}$$

Therefore $\mathbb{P}[X \geqslant a] = \mathbb{P}\left[e^{tX} \geqslant e^{ta}\right] \underbrace{\leqslant}_{\text{Markov}} \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$.

$$\mathbb{P}[X \geqslant (1+\delta)\mu] = \mathbb{P}\left[e^{tX} \geqslant e^{t(1+\delta)\mu}\right] \underbrace{\leqslant}_{\text{Markov}} \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\delta)\mu}} \qquad (*)$$

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^{t(\sum_{i=1}^{n} X_i)}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] \underbrace{=}_{\text{Ind.}X_i} \prod_{i=1}^{n} \mathbb{E}\left[e^{tX_i}\right].$$

$$\mathbb{E}\left[e^{tX_i}\right] = p_i e^t + (1-p_i)e^0 = p_i(e^t - 1) + 1 < e^{p_i(e^t-1)}.$$

$$\therefore \prod_{i=1}^{n} \mathbb{E}\left[e^{tX_i}\right] < \prod_{i=1}^{n} e^{p_i(e^t-1)} = e^{\sum_{i=1}^{n} p_i(e^t-1)} \underbrace{=}_{e^t=\Theta(1)} e^{\mu(e^t-1)}.$$

From (*): $\mathbb{P}[X \geqslant (1+\delta)\mu] < \frac{e^{\mu(e^t-1)}}{e^{t(1+\delta)\mu}} = e^{\mu(e^t-1-t-\delta t)}$

Proof of Theorem 1: Upper tail

Proof (cont'd)

We got $\mathbb{P}[X \geqslant (1+\delta)\mu] < e^{\mu(e^t - 1 - t - \delta t)}$.
To get a tight bound we have to choose t s.t. it minimizes the above expression.
i.e. we have to derivate wrt t: $\frac{d}{dt}\left(e^t - 1 - t - \delta t\right) = 0 \Rightarrow t = \ln(\delta + 1)$
Substituting in the above equation:

$$\mathbb{P}[X \geqslant (1+\delta)\mu] < e^{\mu((\delta+1) - 1 - \ln(\delta+1) - \delta \ln(\delta+1))}$$

$$= \left(\frac{e^{\delta+1-1}}{e^{(\delta+1)\ln(\delta+1)}}\right)^{\mu} = \left(\frac{e^{\delta}}{(\delta+1)^{\delta+1}}\right)^{\mu}.$$

$\square$

Proof of Theorem 1: Lower tail

Proof

As before, we write inequality as inequality in exponents, multiplied by a $t > 0$, which we minimized to get the sharp bound.

We use Markov, but the inequality would be reversed:

$\mathbb{P}[X < (1-\delta)\mu] = \mathbb{P}\left[e^{-tX} > e^{-t(1-\delta)\mu}\right] \leqslant \frac{\mathbb{E}[e^{-tX}]}{e^{-t(1-\delta)\mu}}$.

As $X = \sum X_i$, where $\{X_i\}$ are independent, then $e^{-tX} = \prod_{i=1}^n e^{-tX_i}$,

$\Rightarrow \mathbb{E}\left[e^{-tX}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{-tX_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{-tX_i}\right]$.

But $\mathbb{E}\left[e^{-tX_i}\right] = p_i e^{-t} + (1-p_i)e^0 = p_i e^{-t} + (1-p_i) = 1 - p_i(1 - e^{-t}) \underbrace{\leqslant}_{e^{-t} \geqslant 1-t} e^{-p_i(1-e^{-t})} \leqslant e^{p_i(e^{-t}-1)}$

$\Rightarrow \prod_{i=1}^n \mathbb{E}\left[e^{-tX_i}\right] < \prod_{i=1}^n e^{p_i(e^{-t}-1)} = e^{\sum_i p_i(e^{-t}-1)} = e^{(\mu(e^{-t}-1))}$

So $\mathbb{P}[X < (1-\delta)\mu] < \frac{e^{(\mu(e^{-t}-1))}}{e^{-t(1-\delta)\mu}} = e^{\mu(e^{-t}+t-t\delta-1)}$.

Proof of Theorem 1: Lower tail

Proof (cont'd)

We have to minimize wrt t: $\mathbb{P}[X < (1-\delta)\mu] < e^{\mu(e^{-t}+t-t\delta-1)}$.

$\frac{d}{dt}\mu\left(e^{-t}+t-t\delta-1\right) = 0 \Rightarrow t = \ln\frac{1}{1-\delta}$.

Substituting back into the above equation,

$$\begin{aligned}
\mathbb{P}[X < (1-\delta)\mu] &< e^{\mu((-e^{\ln(1/(1-\delta))})+(1-\delta)\ln(1/(1-\delta))-1)} \\
&= e^{\mu((1-\delta)+(1-\delta)(\ln(1)-\ln(1-\delta))-1)} \\
&= e^{\mu((1-\delta)-1+1/((1-\delta)^{1-\delta})} = \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}
\end{aligned}$$

$\square$

Powerful Technique: Chernoff + Union-Bound

Assume we have an event $A = \cup_{i=1}^{n} A_i$, where the $\{A_i\}_{i=1}^{n}$ are **not** independent, and we want to prove that the probability that A has a bad instance goes $\to 0$.

The technique consists in:

1. Use Chernoff to prove that for each A_i the probability of a bad instance is very small, for each A_i of the n ones, i.e. we compute that $\mathbb{P}[A_i \text{ is bad}]$ is very small,
2. Use Union-Bound to prove $\mathbb{P}[A \text{ is bad}] = \mathbb{P}\left[\cup_{i=1}^{n} A_i \text{ is bad}\right] \leqslant \sum_{i=1}^{n} \mathbb{P}[A_i \text{ is bad}]$ is very small.

Notice, that means that we need $\mathbb{P}[A_i \text{ is bad}] = o(1/n)$, so the sum does not affect $\mathbb{P}[A \text{ is bad}]$.

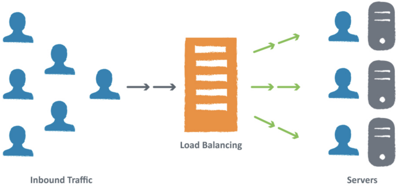
Load balancing problem

Example

Suppose we have k servers and n jobs, $n \gg k$. Assume n jobs stream sequentially but very quickly, we have to assign each job to a server, where each job take a while to process. We are interested in to keep similar load in each servers. We want to have an algorithm that on the fly distribute the jobs into the servers, to balance the load between them, as much as we can.

Random algorithm for load balancing

We want to see "how close" our the load balance achieved by our algorithm is to the perfect load balance $= n/k$,
i.e. we prove that w.h.p., the maximum load of all the servers is near n/k



Randomized Algorithm: Assign independently each job to a random server, with probability $= 1/k$.

Load balancing: correctness

For ($1 \leqslant i \leqslant k$) let X_i be a r.v. counting the number of jobs handled by server i (notice these are not indicator r.v.)

For each $X_i \sim \text{Bin}(n, \frac{1}{k}) \Rightarrow \mathbb{E}[X_i] = \frac{n}{k}$

But $(X_1, \ldots, X_k)$ are not independent, as

$$\underbrace{\mathbb{P}[(X_1 = n) \cap \cdots \cap (X_k = n)]}_{=0} \neq (\underbrace{\mathbb{P}[X_1 = n] \cdots \mathbb{P}[X_k = n]}_{=(\frac{1}{k})^{kn}}.$$

Let M be a r.v. counting the maximum load among all the k servers. $M = \max\{X_1, \ldots, X_k\}$

We want to show $\mathbb{P}\left[M \geqslant \frac{n}{k} + \gamma\right]$ very small, for some γ not too large.

Load balancing: correctness

For any $1 \leqslant i \leqslant k$ define the bad event B_i as $B_i \equiv X_i \geqslant \frac{n}{k} + \gamma$,

Define the event $B = \cup_{i=1}^{k} B_i$, i.e B is the event $M \geqslant \frac{n}{k} + \gamma$.

We aim to show that $\mathbb{P}[B] \leqslant \frac{1}{k^2}$, $\Rightarrow \mathbb{P}\left[\bar{B}\right] > 1 - \frac{1}{k^2}$.

Notice that for all $1 \leqslant i \leqslant k$ we have the same value of $\mathbb{P}[B_i]$. therefore, let $\mathbb{P}[B_i] = \mathbb{P}\left[X_i \geqslant \frac{n}{k} + \gamma\right] = \beta$.

To get $\mathbb{P}[B] \leqslant \frac{1}{k^2}$, using Union Bound:
$\mathbb{P}[B] \leqslant \sum_{i=1}^{k} \mathbb{P}[B_i] = k\beta$, we need $k\beta = \frac{1}{k^2}$, $\Rightarrow$ we need $\mathbb{P}[B_i] = \beta \leqslant \frac{1}{k^3}$.

W.l.o.g. let us compute $\mathbb{P}[B_1]$.

Upper bound for $\mathbb{P}[B_1]$

As $X_1 \sim \text{Bin}(n, \frac{1}{k})$, then $X_1 = \sum_{j=1}^{n} I_j$, where I_j is the indicator r.v. that is 1 if job j goes to server 1. So $\mathbb{P}\left[I_j = 1\right] = \frac{1}{k}$.
$\Rightarrow \mathbb{E}[X_1] = \mu = \sum_{i=1}^{n} \frac{1}{k} = \frac{n}{k}$.

We use Cor. 2.2 to bound $\mathbb{P}[B_1] = \mathbb{P}[X_1 \geqslant \mu + \gamma]$.

$$\mathbb{P}\left[X_1 \geqslant (1+\delta)(\tfrac{n}{k})\right] = \mathbb{P}\left[X_1 \geqslant \left(\frac{n}{k} + \underbrace{\frac{\delta n}{k}}_{\gamma}\right)\right] \leqslant e^{-\frac{\delta^2 \mu}{3}}$$

We need to find suitable values of δ and γ to make everything work.

Choosing values of δ and γ

We know $n \gg k$, we want $\delta < 1$ and $\mathbb{P}[B_1] \leqslant 1/k^3$, then we can make

$$\frac{1}{k^3} = e^{-\frac{\mu\delta^2}{3}}.$$

Taking logarithms in both sides:
$\mu\delta^2 = 9 \ln k \Rightarrow \delta = 3\sqrt{\ln k}\sqrt{k/n}$.

As $\gamma = \frac{\delta n}{k} \Rightarrow \gamma = 3\sqrt{\ln k}\sqrt{n/k}$.

Therefore, $\mathbb{P}[B_1] = \mathbb{P}\left[X_1 \geqslant \mu + 3\sqrt{\frac{n \ln k}{k}}\right] \leqslant \frac{1}{k^3}$,

and $\mathbb{P}[B] \leqslant \frac{1}{k^2}$.

The final result

We have proved that the simple randomized algorithm to allocate n jobs to k servers, with $n \geqslant 9k \ln k$, we get that the algorithm produces a good load balancing, where the probability of having a bad event, is $\leqslant 1/k^3$, i.e., a bad event is that the loads in one server deviates more that $3\sqrt{\ln k}\sqrt{n/k}$ from the expected load n/k.

Therefore. w.h.p. the randomized algorithm will keep the load concentrated around n/k.

Consequences

In practice, how good is that bound ?

Pretty good! If $n = 10^6$ and $k = 10^3$, $n/k = 10^3$ and $\gamma = 250$.
So the result $\Rightarrow$ w.h.p. , the maximum load is $\leqslant$ 1250.

There are better algorithms to the load distribution's problem, but they use more advanced probability techniques, as the power of two choices.

Chernoff: More Sampling

(See also section 4.2.3 in MU book)

We want to poll a sample of size n from a large population of N individuals, about the if they like or they do not like, a given product (answer yes/no).

We want to estimate the real fraction p ($0 < p < 1$) of the population N, that likes the product, i.e. $p =$ #yes votes$/N$.

For that, we sample u.a.r. n persons, i.e. with replacement, and want to know how large n should be so the sampling yields an estimation $\tilde{p} =$ #yes answers$/n$ of the likeness of the product, which is "accurate" and has a high "confidence".

Sampling: Accuracy and confidence

- Accuracy: It is difficult to pinpoint exactly the value of p, so we consider a $\delta > 0$ (the accuracy), and define an interval $[\tilde{p} - \delta, \tilde{p} + \delta]$, such that $\mathbb{P}[p \in [\tilde{p} - \delta, \tilde{p} + \delta]]$ is very high.
- Confidence: choosing γ as small as possible so that $\mathbb{P}[p \in [\tilde{p} - \delta, \tilde{p} + \delta]] \geqslant 1 - \gamma$, where $1 - \gamma$ is the confidence.

Notice we have to tune the values of n, δ and γ as to optimize the accuracy δ with as high as possible confidence $1 - \gamma$.

In a poll, we want to be able to say things like:
This poll is 3% accurate, with 95% confidence
which means that the confidence is $1 - \gamma = 0.95$, the outcome on the whole population N is $\pm 3\%$ of our obtained prediction $\tilde{p}$, i.e. the accuracy is $\delta = 0.03$.

Sampling

Let n be the selected number of people that we poll. Define a set of independent r.v. $\{X_i\}_{i=1}^n$, where each $X_i = 1$ if the i-th person would vote for the product, otherwise $X_i = 0$.

Let $X = \sum_{i=1}^n X_i$, then $X \sim \text{Bin}(n, \tilde{p})$ and X counts the number of people who likes the product.

We want to compute how large do we have to make n to have a good accuracy δ with high confidence $1 - \gamma$.

Sampling Theorem

Theorem (Sampling Theorem)

Suppose we use independent, uniformly random samples (with replacement) to compute an estimate $\tilde{p}$, for p. If the number n of samples satisfies $n \geqslant \frac{3}{\delta^2} \ln \frac{2}{\gamma}$, then

$$\mathbb{P}[p \in [\tilde{p} - \delta, \tilde{p} + \delta]] \geqslant 1 - \gamma.$$

Proof of the sampling theorem

Proof

Given a particular sampling of n people, we find that exactly $n\tilde{p}$ people like the product. we have to find values of δ and γ s.t.:

$$\mathbb{P}[p \in [\tilde{p} - \delta, \tilde{p} + \delta]] = \mathbb{P}[np \in [n(\tilde{p} - \delta), n(\tilde{p} + \delta)]] \geqslant 1 - \gamma.$$

If $p \notin [\tilde{p} - \delta, \tilde{p} + \delta]$ is because either,

- $p < \tilde{p} - \delta \Rightarrow X = n\tilde{p} > n(p + \delta) = \mu(1 + \delta/p)$, or
- $p > \tilde{p} + \delta \Rightarrow X = n\tilde{p} < n(p - \delta) = \mu(1 - \delta/p)$.

Using Corollary 2

$$\begin{aligned}\mathbb{P}[p \notin [\tilde{p} - \delta, \tilde{p} + \delta]] &= \mathbb{P}[X < np\,(1 - \delta/p)] + \mathbb{P}[X > np\,(1 + \delta/p)] \\ &< e^{-n\delta^2/2p} + e^{-n\delta^2/3p}\end{aligned}$$

As $p \leqslant 1$ we get

$$\mathbb{P}[p \in [\tilde{p} - \delta, \tilde{p} + \delta]] = 1 - \mathbb{P}[p \notin [\tilde{p} - \delta, \tilde{p} + \delta]] \geqslant 1 - 2e^{-n\delta^2/3}.$$

But if we want confidence $1 - \gamma$, then we need $\gamma \geqslant 2e^{-\frac{n\delta^2}{3}}$
$\Rightarrow \frac{2}{\gamma} \leqslant e^{\frac{n\delta^2}{3}} \Rightarrow \frac{2}{\gamma} \leqslant \frac{n\delta^2}{3} \Rightarrow n \geqslant \frac{3}{\delta^2} \ln \frac{2}{\gamma}$ $\square$

Sampling Theorem: Some comments

Example

In the previous example, $\delta = 3\%$ and confidence 95% i.e., $\gamma = 1/20$, then we need $n \geqslant \lceil \frac{3}{0.02^2} \ln \frac{2}{1/20} \rceil = 12297$ people giving valid answers.

- Notice in the Sampling Theorem, the number of samples n does not depend on the size N of the total population, i.e., the number of samples you need to get a certain accuracy and a certain confidence only depends on that accuracy and confidence.
- Computing a high accuracy could be costly in the number n of samples, because of the $1/\delta^2$ term. We should design the sampling to tune between accuracy and a realistic sampling of people.
- Getting really high confidence is cheap: because of the logarithm, it hardly costs anything to get a very small γ.