More concentration bounds

AGT-MIRI QT 2020-2021

Why we need more concentration bounds?

- Remember that given a random variable, we are trying to determine how concentrate it is, i.e. that 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 $X \geq 0$ Markov can give an indication that there are values very far away from its mean, but in general is to weak for proving strong concentration results.
- Chebyshev's inequality can give stronger results for concentration of X around μ, but we must compute $\mathbf{Var}[X]$, which could be difficult.

Chernoff Bounds

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 Poisson trials, where each X_i can has probability of success p_i. The particular case where all p_i are equal is the Bernouilli trials.

Theorem (Ch-1) Let $\{X_i\}_{i=0}^n$ be independent Poisson trials, with $\mathbf{Pr}[X_i = 1] = p_i$. Then, if $X = \sum_{i=1}^n X_i$, and $\mu = \mathbf{E}[X]$, we have

1. $\mathbf{Pr}[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$, for $\delta \in (0,1)$.
2. $\mathbf{Pr}[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$ for any $\delta > 0$.

Weak Chernoff's bound, but easy to use

Corollary (Ch-2) Let $\{X_i\}_{i=0}^n$ be independent Poisson trials, with $\mathbf{Pr}[X_i = 1] = p_i$. Then if $X = \sum_{i=1}^n X_i$, and $\mu = \mathbf{E}[X]$, we have

1. $\mathbf{Pr}[X \leq (1-\delta)\mu] \leq e^{-\mu\delta^2/2}$, for $\delta \in (0,1)$.
2. $\mathbf{Pr}[X \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/3}$, for $\delta \in (0,1]$.

An immediate corollary to the previous result:

Corollary (Ch-3) Let $\{X_i\}_{i=0}^n$ be independent Poisson trials, with $\mathbf{Pr}[X_i = 1] = p_i$. Then if $X = \sum_{i=1}^n X_i$, $\mu = \mathbf{E}[X]$ and $\delta \in (0,1)$, we have

$$\mathbf{Pr}[|X - \mu| \geq \delta\mu] \leq 2e^{-\mu\delta^2/3}.$$

Sketch Proof of (3) using (2):

$$\begin{aligned}\mathbf{Pr}[|X-\mu| \geq \delta\mu] &= \mathbf{Pr}[X < (1-\delta)\mu] + \mathbf{Pr}[X \geq (1+\delta)\mu] \\ &\leq e^{-\mu\delta^2/2} + e^{-\mu\delta^2/3} \leq 2e^{-\mu\delta^2/3}\end{aligned}$$

Proof of Ch-2.1 using Ch-1.1

From (Ch-2.1) We must prove that, for $\delta \in (0,1)$, we have $\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \leq e^{-\mu\delta^2/2} = \left(e^{-\delta^2/2}\right)^{\mu}$.

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

$f(\delta) = -\delta - (1-\delta)\ln(1-\delta) + \delta^2/2 \leq 0.$

Differenciating $f(\delta)$:

$$f'(\delta) = \ln(1-\delta) + \delta$$

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

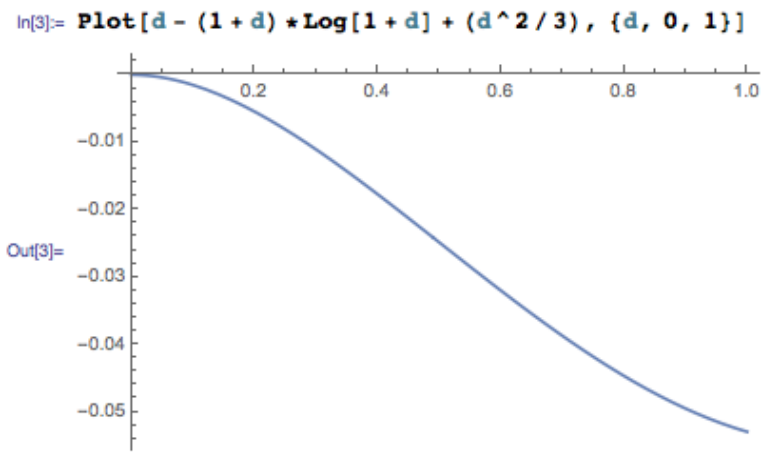
$\Rightarrow f''(\delta) < 0$ in $(0,1)$ and as $f'(0) = 0$, then $f'(\delta) \leq 0$ in $[0,1)$, i.e. $f(\delta)$ is non-increasing in $[0,1)$.

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

Proof of Ch-2.2 using Ch-1.2

From (Ch-2.2) We must prove that for $\delta \in (0,1)$, we have $\left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \leq e^{-\delta^2/3}$.

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



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 $X \in B(n, 1/2)$, then, $\mu = n/2, \mathbf{Var}[X] = n/4$.

We want to bound $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right]$.

- Markov: $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq \frac{\mu}{3n/4} = 2/3$.
- Chebyshev's: $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq \mathbf{Pr}\left[|X - \frac{n}{2}| \geq \frac{n}{4}\right] \leq \frac{\mathbf{Var}[X]}{(n/4)^2} = \frac{4}{n}$.
- Chernoff: We want $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right]$. Using Ch-2.2,
  $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right] = \mathbf{Pr}\left[X \geq (1+\delta)\frac{n}{2}\right] \Rightarrow (1+\delta)\frac{3}{2} \Rightarrow \delta = \frac{1}{2}$
  $\therefore \mathbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq e^{-\mu\delta^2/3} = e^{-\frac{n}{24}}$.

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

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

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

1. Markov: $\mathbf{Pr}[X \geq n/2] \leq \frac{n/2}{n/2} = 1$. So $\mathbf{Pr}[X \leq n/2] \geq 0$. No Information
2. Chebyshev: Between $n/4$ and $3n/4$ heads:
   $\mathbf{Pr}\left[|X - \frac{n}{2}| \geq \frac{n}{4}\right] \leq \frac{4}{n}$
3. Chernoff: Using the last bound
   $\mathbf{Pr}\left[|X - \frac{n}{2}| \geq \frac{1}{2}\sqrt{6n \ln n}\right] \leq 2e^{-\frac{1}{3}\frac{n}{2}\frac{6 \ln n}{n}} = \frac{2}{n}$
   Even $\mathbf{Pr}\left[|X - \frac{n}{2}| \geq \frac{n}{4}\right] \leq 2e^{-\frac{1}{3}\frac{n}{2}\frac{1}{4}} \leq 2e^{-\frac{n}{24}}$

Proof of Chernoff-1: Upper tail

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

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

Therefore $\mathbf{Pr}[X \geq a] = \mathbf{Pr}\left[e^{tX} \geq e^{ta}\right] \underbrace{\leq}_{\text{Markov}} \frac{\mathbf{E}\left[e^{tX}\right]}{e^{ta}}$.

$$\mathbf{Pr}[X \geq (1+\delta)\mu] = \mathbf{Pr}\left[e^{tX} \geq e^{t(1+\delta)\mu}\right] \underbrace{\leq}_{\text{Markov}} \frac{\mathbf{E}\left[e^{tX}\right]}{e^{t(1+\delta)\mu}} \qquad (*)$$

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

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

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

From (*): $\mathbf{Pr}[X \geq (1+\delta)\mu] < \frac{e^{\mu(e^t-1)}}{e^{t(1+\delta)\mu}} = e^{\mu(e^t-1-t-\delta t)}$

Proof of Chernoff-1: Upper tail

We got $\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] < e^{\mu(e^t-1-t-\delta t)}$.
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: $\frac{\mathrm{d}(e^t-1-t-\delta t)}{\mathrm{d}t} = 0 \Rightarrow t = \ln(\delta+1)$
Substituting in the above equation:

$$\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] < e^{\mu((\delta+1)-1-\ln(\delta+1)-\delta\ln(\delta+1))}$$

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

Proof of Chernoff-2: Lower tail

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

As before we use Markov, but the inequality would be reversed:

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

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

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

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

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

So $\mathbf{Pr}\left[X < (1-\delta)\mu\right] < \frac{e^{(\mu(e^{-t}-1))}}{e^{-t(1-\delta)\mu}} = e^{\mu(e^{-t}+t-t\delta-1)}$.

Proof of Chernoff-2: Lower tail

We have to minimize wrt t: $\mathbf{Pr}\left[X < (1-\delta)\mu\right] < e^{\mu(e^{-t}+t-t\delta-1)}$.
$\frac{\mathrm{d}\mu(e^{-t}+t-t\delta-1)}{\mathrm{d}t} = 0 \Rightarrow t = \ln\frac{1}{1-\delta}$.
Substituting back into the above equation,

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

□

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 $\rightarrow 0$ (it is tiny).
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 $\mathbf{Pr}[A_i \text{ is bad}]$ is very small,
2. use Union-Bound to prove $\mathbf{Pr}[A \text{ is bad}] = \mathbf{Pr}[\cup_{i=1}^{n} A_i \text{ is bad}] \leq \sum_{i=1}^{n} \mathbf{Pr}[A_i \text{ is bad}]$ is very small.

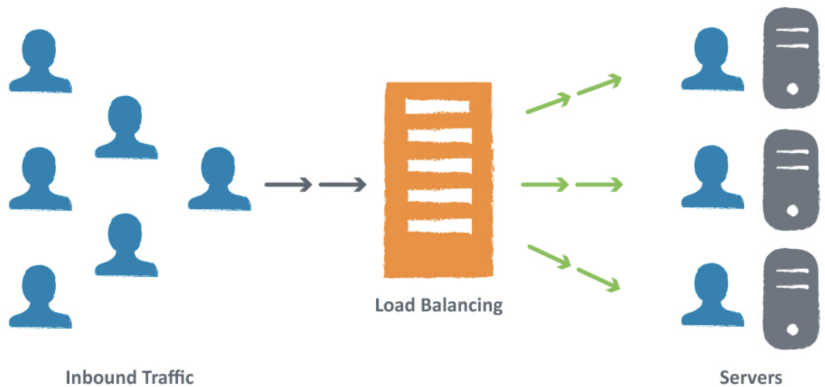
Notice, that means that we need $\mathbf{Pr}[A_i \text{ is bad}] = \omega(1/n)$, so the sum does not affect $\mathbf{Pr}[A \text{ is bad}]$.

Load balancing problem

Suppose we have k servers and n jobs, $n >> k$. 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. 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 $(1 \leq i \leq k)$ let X_i be a r.v. counting the number of jobs handled by server i. (notice they are not indicator r.v.)

For each $X_i \in B(n, \frac{1}{k}) \Rightarrow \mathbf{E}[X_i] = \frac{n}{k}$ clear?

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

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

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 $\mathbf{Pr}\left[M \geq \frac{n}{k} + \gamma\right]$ very small, for a not too large γ.

Correctness-2

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

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

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

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

To get $\mathbf{Pr}[B] \leq \frac{1}{k^2}$, using Union Bound:
$\mathbf{Pr}[B] \leq \sum_{i=1}^{k} \mathbf{Pr}[B_i] = k\beta$, which we need $= \frac{1}{k^2}$, $\Rightarrow$ we need
$\mathbf{Pr}[B_1] = \beta \leq \frac{1}{k^3}$.

W.o.l.g let us compute $\mathbf{Pr}[B_1]$.

Showing that $\mathbf{Pr}[B_1] \leq 1/k^3$

As $X_1 \in B(n, \frac{1}{k})$, then $X_1 = \sum_{j=1}^{n} I_j$, where I_j is i.r.v. that is 1 if job j goes to server 1. So $\mathbf{Pr}[I_j = 1] = \frac{1}{k}$.
$\Rightarrow \mathbf{E}[X_1] = \mu = \sum_{i=1}^{n} \frac{1}{k} = \frac{n}{k}$.

We use Ch-2.2 to bound $\mathbf{Pr}[B_1] = \mathbf{Pr}[X_1 \geq \mu + \gamma]$.

$$\mathbf{Pr}\left[X_1 \geq (1+\delta)(\tfrac{n}{k})\right] = \mathbf{Pr}\left[X_1 \geq \left(\frac{n}{k} + \underbrace{\frac{\delta n}{k}}_{\gamma}\right)\right] \leq e^{-\frac{\delta^2 \mu}{3}}$$

We need to take values of δ and γ, to make everything work.

Choosing values of δ and γ so everything works

We know $n >> k$, we want $\delta < 1$ and $\mathbf{Pr}[B_1] \leq 1/k^3$, then we can make

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

Taking ln 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, $\mathbf{Pr}[B_1] = \mathbf{Pr}\left[X_1 \geq \mu + 3\sqrt{\frac{n \ln k}{k}}\right] \leq \frac{1}{k^3}$,

and $\mathbf{Pr}[B] \leq \frac{1}{k^2}$. □

The final result

We have proved that the simple randomized algorithm to allocate n jobs to k servers, with $n \geq 9k \ln k$, we get that the algorithm produces a load balancing, where the probability of having a bad event, is $\leq 1/k^3$, 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 ≤ 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. $p = \#\text{yes votes}/N$.

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 $\tilde{p} = \#\text{yes answers}/n$ 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 $\mathbf{Pr}[p \in [\tilde{p} - \delta, \tilde{p} + \delta]]$ is very high.
- Confidence: choosing γ as small as possible so that $\mathbf{Pr}[p \in [\tilde{p} - \delta, \tilde{p} + \delta]] \geq 1 - \gamma$, 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, 19 times out of 20.
Which mean that with confidence $1 - \gamma = 19/20 = 95\%$, 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 $X \in B(n, \tilde{p})$ and X count the number of people that likes the product

Define our "guess" $\tilde{p}$ as $X = \tilde{p}n$.

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

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

$$\mathbf{Pr}\left[p \in [\tilde{p} - \delta, \tilde{p} + \delta]\right] \geq 1 - \gamma.$$

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

$$\mathbf{Pr}\left[p \in [\tilde{p} - \delta, \tilde{p} + \delta]\right] = \mathbf{Pr}\left[np \in [n(\tilde{p} - \delta), n(\tilde{p} + \delta)]\right] \geq 1 - \gamma.$$

Proof of the sampling theorem

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 the Corollary to Ch-2, we get

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

As $p \leq 1$ we get

$$\mathbf{Pr}\left[p \in [\tilde{p} - \delta, \tilde{p} + \delta]\right] = 1 - \mathbf{Pr}\left[p \notin [\tilde{p} - \delta, \tilde{p} + \delta]\right] \geq 1 - 2e^{-n\delta^2/3}.$$

But if we want confidence $1 - \gamma$, then we need $\gamma \geq 2e^{-\frac{n\delta^2}{3}}$

$\Rightarrow \frac{2}{\gamma} \leq e^{\frac{n\delta^2}{3}} \Rightarrow \frac{2}{\gamma} \leq \frac{n\delta^2}{3} \Rightarrow n \geq \frac{3}{\delta^2} \ln \frac{2}{\gamma}$ □

Sampling Theorem: Some comments

In the previous example, $\delta = 3\%$ and confidence 95% i.e. $\gamma = 1/20$, then we need $n \geq \lceil \frac{3}{0.02^2} \ln \frac{2}{1/20} \rceil = 12297$ 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 ln, it hardly costs anything to get a very small δ.