Analysis of Min Count

Last time

- $\mathcal{M} = (S, m)$, $S = \{s_1, \dots s_n\}$, $n = |S| = ?$
- MinCount: $\mathcal{M} \to h(s_i) = U_i \sim U(0,1) \to U_1, U_2, \dots U_n \to$

$$\to U_{1:n} \leq U_{2:n} \leq \dots \leq U_{k:n} \to \hat{n}_k = \frac{k-1}{U_{k:n}}$$

estimator

- $U_{k:n} \sim Beta(k, n+1-k)$, $f_k(x) = \dfrac{x^{k-1}(1-x)^{n-k}}{B(k,n+1-k)}$ for $x \in (0,1)$, where

$$B(k, n+1-k) = \frac{\Gamma(k)\,\Gamma(n+1-k)}{\Gamma(n+1)} \underset{n,k \in \mathbb{N}}{=} \frac{(k-1)!\,(n-k)!}{n!}$$

Estimator Construction

$k=1$: $U_{k:n} \sim Beta(1,n)$, $f_1(x) = n(1-x)^{n-1}$

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

- $E[U_{1:n}] = \int_0^1 x \cdot f_1(x)\,dx = n\int_0^1 x(1-x)^{n-1} = n \cdot B(2,n) =$

$$= n \cdot \frac{\Gamma(2)\,\Gamma(n)}{\Gamma(n+2)} = n \cdot \frac{1!\,(n-1)!}{(n+1)!} = \frac{1}{n+1}$$

then can we use it to estimate $\hat{n} \overset{?}{=} \frac{1}{U_{1:n}} \to E[\hat{n}] = E\left[\frac{1}{U_{1:n}}\right] =$

$$= \int_0^1 \frac{1}{x} f_1(x)\,dx = \int_0^1 \frac{1}{x} n(1-x)^{n-1}dx = \infty$$

x is close to 0 but the integral disconverge

generally $E[g(x)] = \int_0^1 g(x) f(x)\,dx =$

$X \sim f(x)$

now let's try for $k \geq 2$

- $E\left[\frac{1}{U_{k:n}}\right] = \int_0^1 \frac{1}{x} f_k(x)\,dx = \int_0^1 \frac{1}{x} \cdot \frac{x^{k-1}(1-x)^{n-k}}{B(k, n+1-k)}\,dx = \frac{1}{B(k,n+1-k)} \int_0^1 x^{k-2}(1-x)^{n-k} dx$

$\overset{\text{Ex 15}}{=} \dots = \frac{n}{k-1}$

- then we can construct estimator $\hat{n}_k := \frac{k-1}{U_{k:n}} \rightarrow E[\hat{n}_k] = n$

Variance

now we want to calculate variance but

$Var[X] = E[X^2] - E[X]^2$

$\hookrightarrow$ but then $g(x) = \frac{1}{x^2}$ and integral still doesn't converge. So we do it for $k \geq 3$

$$Var[\hat{n}_k] = \frac{n(n-k+1)}{k-2}$$

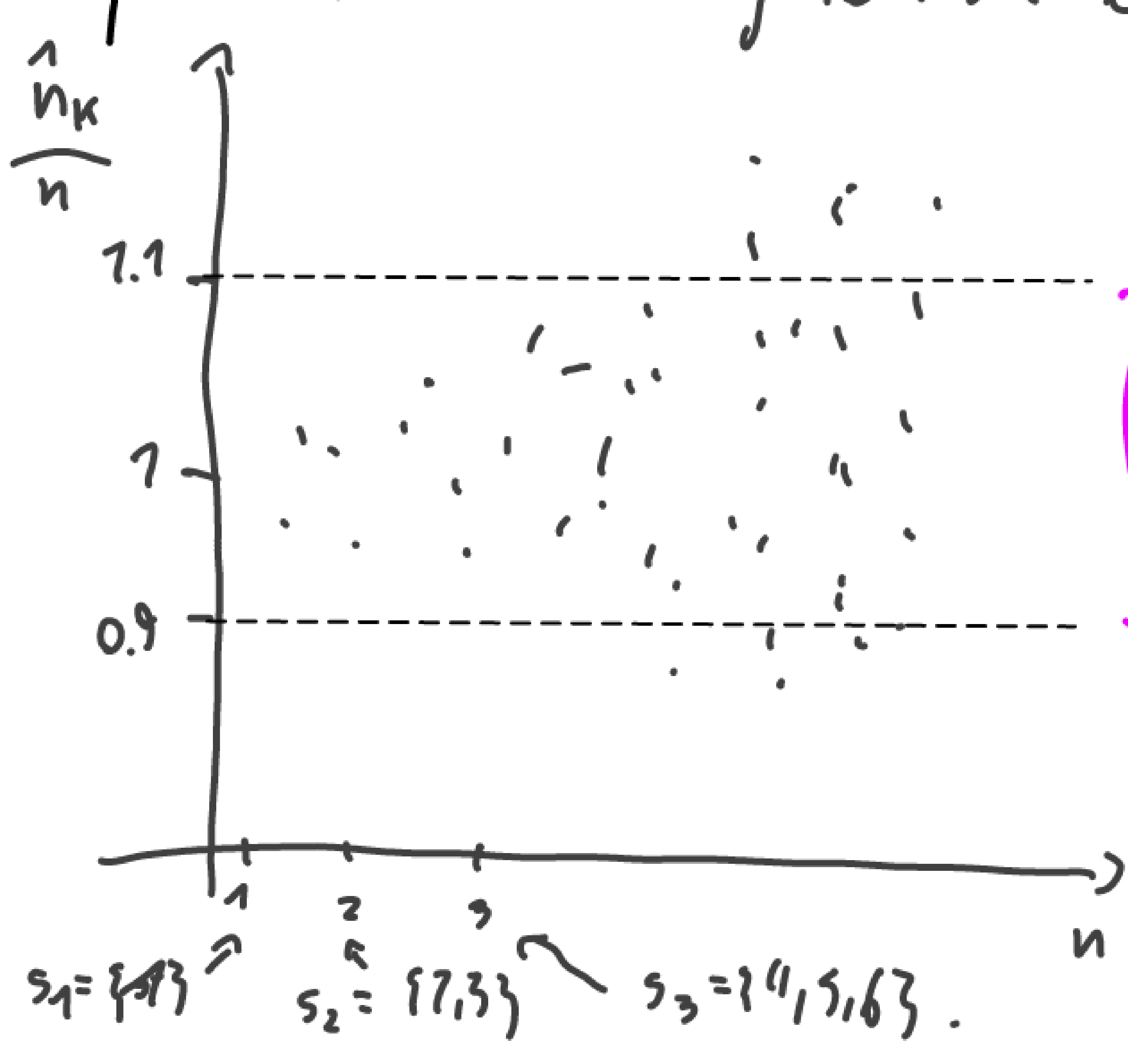
Standard Error

$SE[\hat{n}_k] = \sqrt{Var\left[\frac{\hat{n}_k}{n}\right]}$ ← this can be interpreted as average error expressed in percente

$\approx \frac{1}{\sqrt{k-2}}$, $k = 100 \rightarrow SE[\hat{n}_k] \approx 10\%$

Experiment

we try to make experiment independent

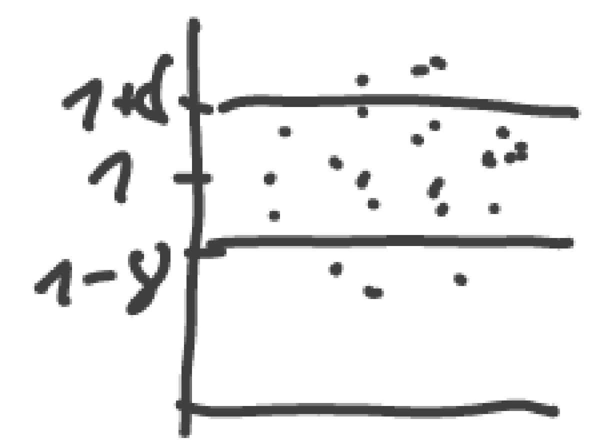


with high probability we would like to sort that 99% of points are in

SE gives only average error and we would like to know

$$Pr\left[\left|\frac{\hat{n}_K}{n}-1\right| > \delta\right] < \alpha \Leftrightarrow Pr\left[\left|\frac{\hat{n}_K}{n}-1\right| < \delta\right] > 1-\alpha$$

we will use Chebyshev inequality (Ex. 17)



Let $\delta > 0$ and X such that $E[X]$, $Var[X]$ are finite.

Then $Pr[|X - EX| < \delta] > 1 - \frac{Var[X]}{\delta^2}$

Example $X = \frac{\hat{n}_K}{n}$, $E[X] = 1$, $\delta = 10\%$
$K = 100$

$$Pr\left[\left|\frac{\hat{n}_K}{n}-1\right| < 10\%\right] > 1 - \frac{Var\left[\frac{\hat{n}_K}{n}\right]}{(10\%)^2} \approx$$

$$1 - \frac{\frac{1}{K}}{\frac{1}{100}} = 0$$, this information is useless.

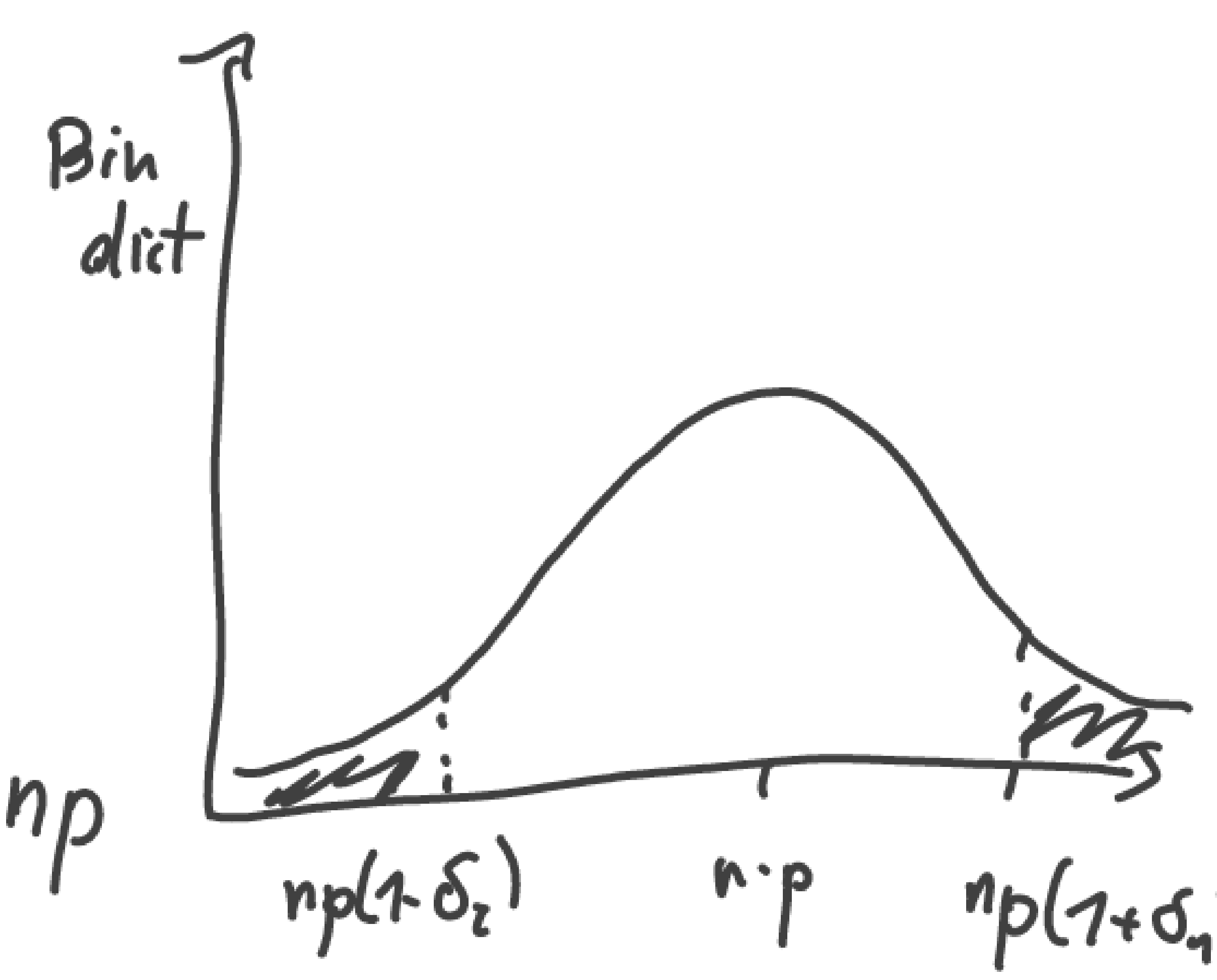
Chernoff inequality (Ex. 18)

version for binomial distribution $B_{n,p} \sim Bin(n,p)$ $\quad E[B_{n,p}] = n \cdot p$

- $Pr[B_{n,p} = k] = \binom{n}{k} p^k (1-p)^{n-k}$
- For $\delta_1 > 0$ and $0 < \delta_2 < 1$ we have

$$(*) \quad Pr[B_{n,p} \geq n \cdot p(1+\delta_1)] \leq \left(\frac{e^{\delta_1}}{(1+\delta_1)^{1+\delta_1}}\right)^{np}$$

$$(**) \quad Pr[B_{n,p} \leq n \cdot p(1-\delta_2)] \leq \left(\frac{e^{-\delta_2}}{(1-\delta_2)^{1-\delta_2}}\right)^{np}$$



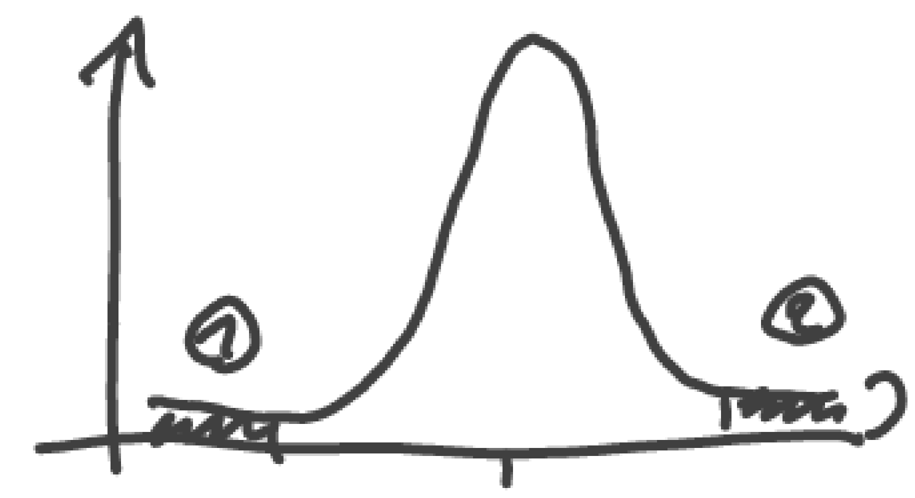
Lemma 1 (Binomial Chernoff → Order Statistic Chernoff)

Let $U_1, U_2, \dots, U_n$ i.i.d., $U_i \sim \mathcal{U}(0,1)$. Let $k \le n$ and $\alpha \in (0,1)$. Then

$$\boxed{E[U_{k:n}] = \frac{k}{n}}$$

1) if $\alpha < \frac{k}{n}$ then $\Pr[U_{k:n} \le \alpha] \le e^{-\alpha n}\left(\frac{\alpha n e}{k}\right)^k$

2) if $\alpha > \frac{k}{n}$ then $\Pr[U_{k:n} \ge \alpha] \le e^{-\alpha n}\left(\frac{\alpha n e}{k}\right)^k$



$E[U_{k:n}] = \frac{k}{n}$

Proof:

- $\mathbb{1}_i = \begin{cases} 1 & \text{if } U_i \le \alpha \\ 0 & \text{if } U_i > \alpha \end{cases}$, $B_{n,\alpha} = \mathbb{1}_1 + \mathbb{1}_2 + \mathbb{1}_3 + \dots + \mathbb{1}_n \sim \mathrm{Bin}(n,\alpha)$

- note that $U_{k:n} \le \alpha \Leftrightarrow |\{i : U_i \le \alpha\}| \ge k \Leftrightarrow B_{n,\alpha} \ge k = \alpha n\left(1 + \frac{k - \alpha n}{\alpha n}\right)$

1) $\alpha < \frac{k}{n}$: $\Pr[U_{k:n} \le \alpha] = \Pr\Big[B_{n,\alpha} \ge \underbrace{\overbrace{\alpha n}^{E[B_{n,\alpha}]}\Big(1 + \overbrace{\frac{k-\alpha n}{\alpha n}}^{\delta}\Big)}_{k}\Big] \overset{(*)}{\le}$

$$\le \left(\frac{e^{\frac{k-\alpha n}{\alpha n}}}{\left(\frac{k}{\alpha n}\right)^{\left(\frac{k}{\alpha n}\right)}}\right)^{n\alpha} = e^{-\alpha n}\left(\frac{\alpha n e}{k}\right)^k$$

2) $\alpha > \frac{k}{n}$: $\Pr[U_{k:n} \ge \alpha] \overset{(**)}{\le} e^{-\alpha n}\left(\frac{\alpha n e}{k}\right)^k$ ∎

Theorem 1 (Chernoff bounds for $\hat{n}_k$)

Let $3 \le k \le n$, $\varepsilon_1 > 0$, $0 < \varepsilon_2 < 1$

Denote $f_k(x) = e^{xk}(1-x)^k$, then for $\hat{n}_k = \frac{k-1}{U_{k:n}}$ we have

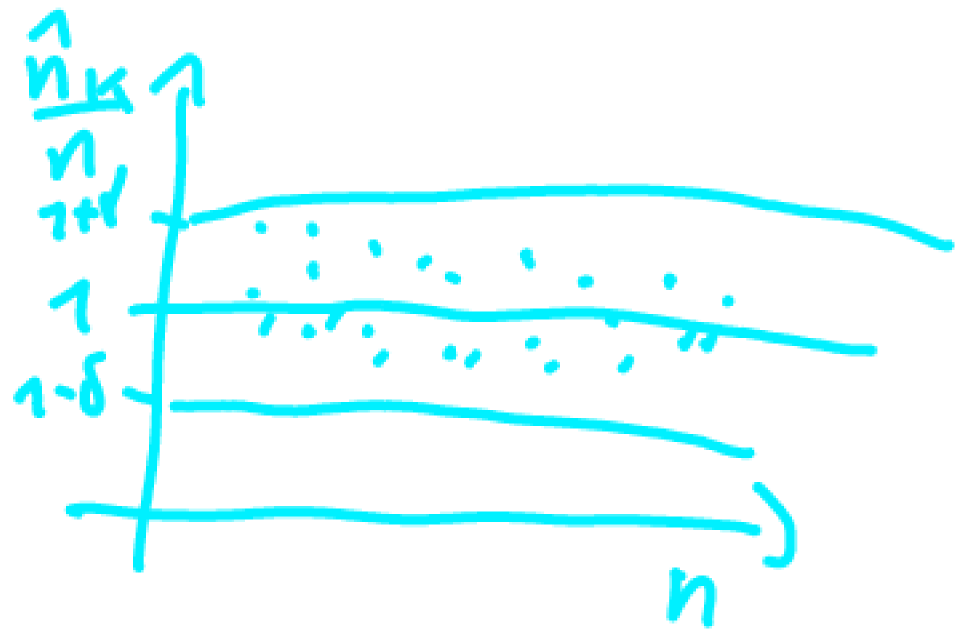
$$\Pr\left[\underbrace{\frac{1}{1+\varepsilon_1}}_{(1-\delta)} \cdot \underbrace{\frac{k-1}{k}}_{\approx 1} < \frac{\hat{n}_k}{n} < \underbrace{\frac{1}{1-\varepsilon_2}}_{(1+\delta)} \cdot \underbrace{\frac{k-1}{k}}_{\approx 1}\right] > 1 - f_k(\overset{\frac{\delta}{1+\delta}}{\varepsilon_2}) - f_k(-\overset{\frac{\delta}{1-\delta}}{\varepsilon_1})$$

$\varepsilon_1 = \frac{\delta}{1-\delta_1}$ $\qquad$ $\varepsilon_2 = \frac{\delta}{1+\delta}$

Notes for Task 7

Look for smallest δ such that $\Pr[1-\delta < \frac{\hat{n}_k}{n} < 1+\delta] > 1-\alpha$, $k = 400$, $\alpha = 1\%$

a) experimentally



b) use Chebyshev inequality

c) use Chernoff (Theorem 1)

Proof: it's obvious

① $\Pr[a < X < b] \overset{(A)}{=} \Pr[X < b] - \Pr[X < a]$

then $\Pr\left[\frac{1}{1+\varepsilon_1}\frac{k-1}{k} < \frac{\hat{n}_k}{n} < \frac{1}{1-\varepsilon_2}\frac{k-1}{k}\right] \overset{(A)}{=} \underbrace{\Pr\left[\frac{\hat{n}_k}{n} < \frac{1}{1-\varepsilon_2}\frac{k-1}{k}\right]}_{\overset{③}{\geq} 1 - f_k(\varepsilon_2)} - \underbrace{\Pr\left[\frac{\hat{n}_k}{n} < \frac{1}{1+\varepsilon_1}\frac{k-1}{k}\right]}_{\overset{②}{\leq} f_k(-\varepsilon_1)}$

② $\Pr\left[\frac{\hat{n}_k}{n} < \frac{1}{1+\varepsilon_1}\frac{k-1}{k}\right] = \Pr\left[\frac{k-1}{U_{k:n}} < \frac{n}{1+\varepsilon_1}\frac{k-1}{k}\right] = \Pr\left[U_{k:n} > \underbrace{\frac{k}{n}(1+\varepsilon_1)}_{\alpha}\right] \overset{\text{Lemma 1}}{\leq} e^{-\alpha n}\left(\frac{\alpha n e}{k}\right)^k$

$= \ldots =$

$= f_k(-\varepsilon_1)$

(3) similar reasoning as (2) $P_n\left[\frac{\hat{n}_K}{n} < \frac{k-1}{K}\frac{1}{1-\varepsilon_2}\right] \geq 1 - f_K(\varepsilon_2) \Leftrightarrow$

$$\leq f_K(\varepsilon_2) \quad \square$$