Performance Evaluation and Networks

Statistics

Statistics & data analysis

Given a dataset from raw observations or from some experimental protocol, statistical methods are used to :

- Clarify/summarize/compress these data in a form that makes their exploitation convenient and efficient (indicators, graphs)
- Model the part of randomness which is underlying in the phenomenon which produced these data (construction of the model by *parameter estimation*, control and validation of the model by *hypothesis testing*).

Vocabulary : population $\supseteq$ sample $\ni$ sample point/unit.
Vocabulaire : population $\supseteq$ échantillon/sondage $\ni$ individu.

Statistics & data analysis

Given a dataset from raw observations or from some experimental protocol, statistical methods are used to :

- Clarify/summarize/compress these data in a form that makes their exploitation convenient and efficient (indicators, graphs) → **descriptive statistics**.
- Model the part of randomness which is underlying in the phenomenon which produced these data (construction of the model by *parameter estimation*, control and validation of the model by *hypothesis testing*). → **inferential statistics**.

Vocabulary : population ⊇ sample ∋ sample point/unit.
Vocabulaire : population ⊇ échantillon/sondage ∋ individu.

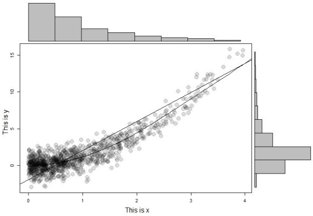
The statistician

- The statistician does not invent his field of investigation, but faces a set of data which, however vast, provides only imperfect knowledge of an underlying reality.
- The statistician does not invent his problem, but he has an interlocutor who expresses, + or - confusingly, expectations regarding the data : clarify/model/predict/decide ...
- A mixture of mathematician, computer scientist, investigator and sometimes specialized in a field of application : economics, social sciences, medicine, ...
- A useful ally at all stages and especially as a last resort : able to make any raw data set talk !
- May be a robot in the near future ...

Graphics

Visualization of samples :

- Use classical charts, e.g., scatter plots for raw data, bars or histograms for distributions, or invent new ones
- Extract/project/mix components if individuals in the sample have many dimensions (e.g., points in $\mathbb{R}^d$)
- Tools available in most stats softwares



Statistical indicators

Indicator : informative numerical value on a sample

- position : central tendency of the sample
- dispersion : deviations from the central value
- shape : asymmetry, flattening of the distribution, hills ...

Two classical categories of indicators : based on ranks (for sorted dataset) or on moments (as defined in proba).

Remark : def for samples can be translated for proba distributions (and vice versa) via the empiral measure assoc to the sample

Definition (Empirical measure/law associated with a sample $x_1, \ldots, x_n$)

discrete distribution $f(x) = \frac{card\{i | x_i = x\}}{n}$ *(link stats ↔ probas)*

Classical indicators of position/dispersion

Two versions :

- statistical : given a sorted sample $x_1 < \cdots < x_n$ of reals
- probabilistic : given a real random variable X (discrete or continuous)

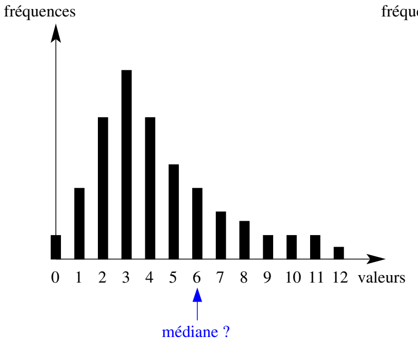
Position	**stats version**	**proba version**
Mean μ	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\mathbb{E}X$
Median m	$x_{[\frac{n+1}{2}]}$	$\mathbb{P}(X < m) \leq \frac{1}{2}$, $\mathbb{P}(X > m) \leq \frac{1}{2}$
Mode M	argmax empirical law	argmax law of X

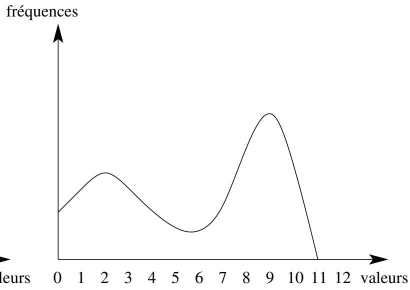
Dispersion	**stats version**	**proba version**
α-quantile q_α	$x_{[\alpha(n+1)]}$	$\mathbb{P}(X < q_\alpha) \leq \alpha$, $\mathbb{P}(X > q_\alpha) \leq 1-\alpha$
Variance σ^2	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$	$\mathbb{E}(X - \mathbb{E}X)^2$

Notation : α-quantile for $0 \leq \alpha \leq 1$ and $[.]$ = choose $\lceil . \rceil$ or $\lfloor . \rfloor$

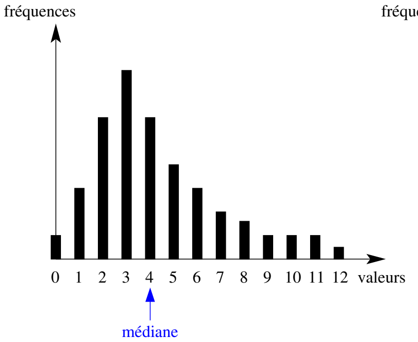
Vocabulary : use “empirical” to qualify stats defs

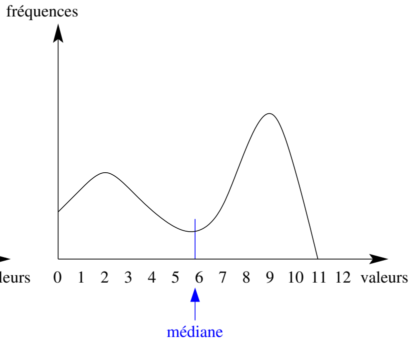
Indicators : boite à moustaches / box plot



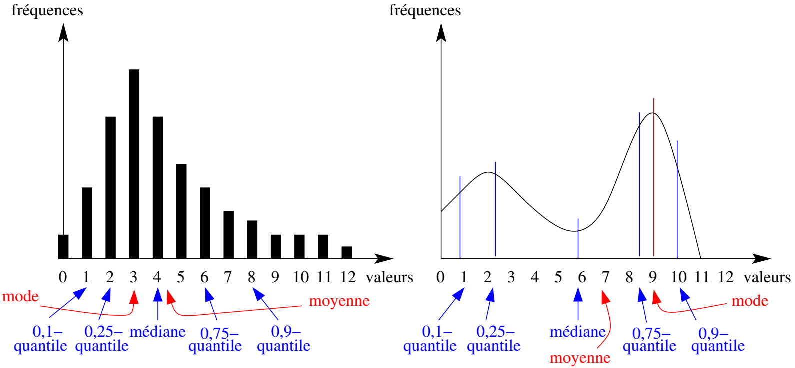


Indicators : boite à moustaches / box plot

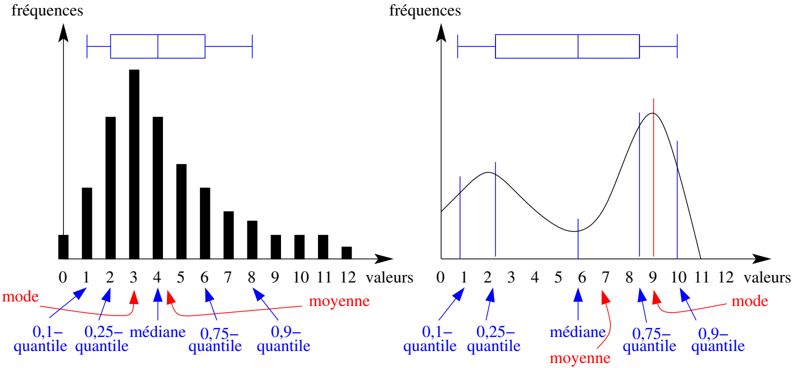




Indicators : boite à moustaches / box plot



Indicators : boite à moustaches / box plot



boite à moustaches / box plot = (0.1-quantile,0.25-quantile,médiane,0.75-quantile,0.9-quantile)

Computation of the classical indicators

Algorithmic complexity for a sample of n unsorted data values :

Mean (empirical)	
Variance (empirical)	
Mode (maximum)	
Median	
α-percentile	
Sorting	

Computation of the classical indicators

Algorithmic complexity for a sample of n unsorted data values :

Mean (empirical)	$\mathcal{O}(n)$
Variance (empirical)	
Mode (maximum)	
Median	
α-percentile	
Sorting	

Computation of the classical indicators

Algorithmic complexity for a sample of n unsorted data values :

Mean (empirical)	$\mathcal{O}(n)$
Variance (empirical)	$\mathcal{O}(n)$
Mode (maximum)	
Median	
α-percentile	
Sorting	

Computation of the classical indicators

Algorithmic complexity for a sample of n unsorted data values :

Mean (empirical)	$\mathcal{O}(n)$
Variance (empirical)	$\mathcal{O}(n)$
Mode (maximum)	$\mathcal{O}(n)$
Median	
α-percentile	
Sorting	

Computation of the classical indicators

Algorithmic complexity for a sample of n unsorted data values :

Mean (empirical)	$\mathcal{O}(n)$
Variance (empirical)	$\mathcal{O}(n)$
Mode (maximum)	$\mathcal{O}(n)$
Median	$\mathcal{O}(n)$
α-percentile	
Sorting	

Computation of the classical indicators

Algorithmic complexity for a sample of n unsorted data values :

Mean (empirical)	$\mathcal{O}(n)$
Variance (empirical)	$\mathcal{O}(n)$
Mode (maximum)	$\mathcal{O}(n)$
Median	$\mathcal{O}(n)$
α-percentile	$\mathcal{O}(n)$
Sorting	

Computation of the classical indicators

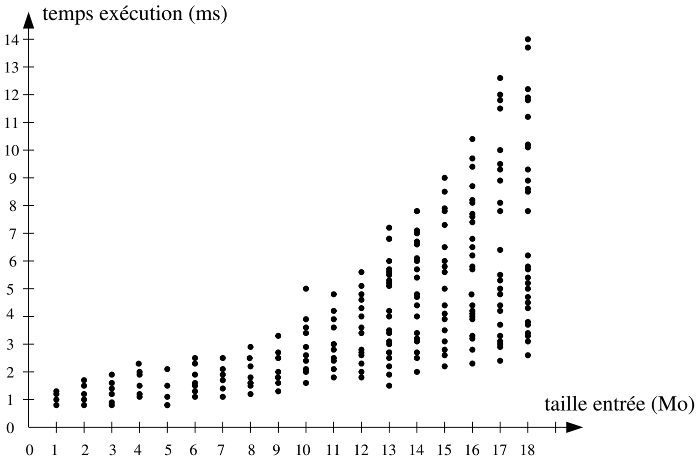
Algorithmic complexity for a sample of n unsorted data values :

Mean (empirical)	$\mathcal{O}(n)$
Variance (empirical)	$\mathcal{O}(n)$
Mode (maximum)	$\mathcal{O}(n)$
Median	$\mathcal{O}(n)$
α-percentile	$\mathcal{O}(n)$
Sorting	from $\mathcal{O}(n)$ to $\mathcal{O}(n \log n)$

Choosing indicators : mode vs mean vs median

	Mean	Median
Algebraic handling	☺	
Use of all data	☺	
Robustness against outliers		☺
Return a value from the dataset		☺

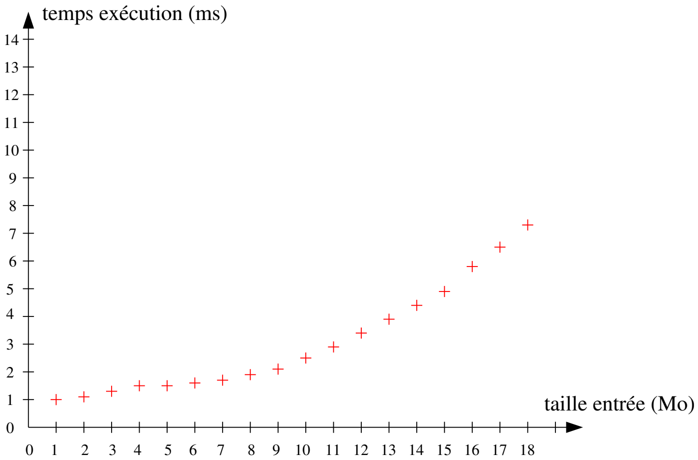
Using indicators : an example



Running time of a software according to input size

100 measures per size

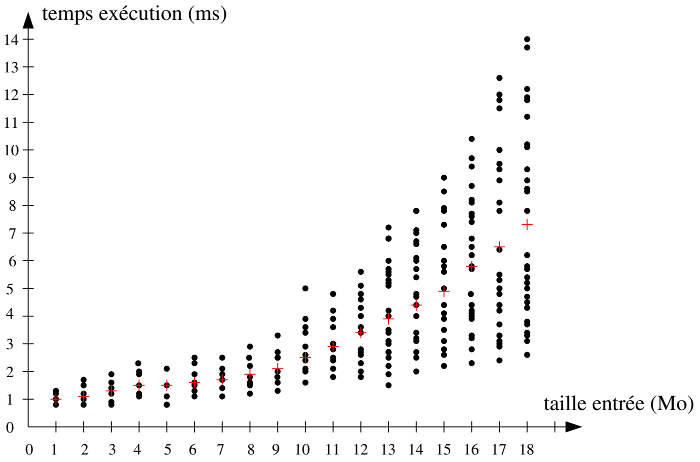
Using indicators : an example



Running time of a software according to input size

100 measures per size   mean per size

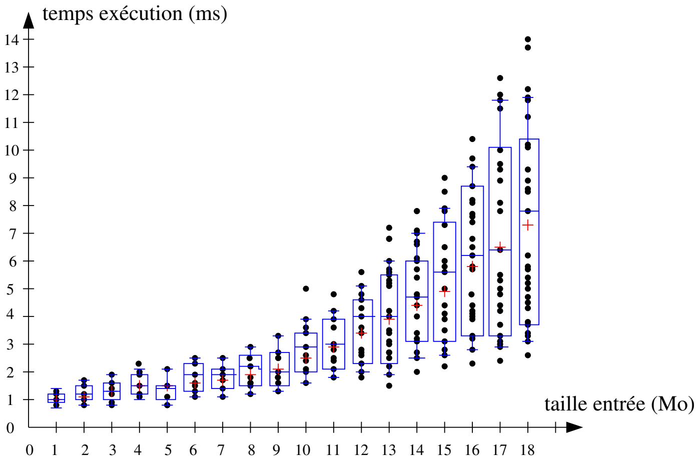
Using indicators : an example



Running time of a software according to input size

100 measures per size mean per size

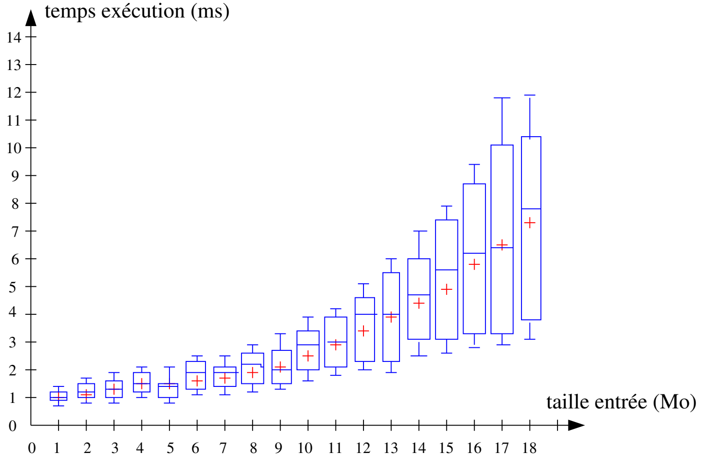
Using indicators : an example



Running time of a software according to input size

100 measures per size   mean per size   boxplot per size

Using indicators : an example



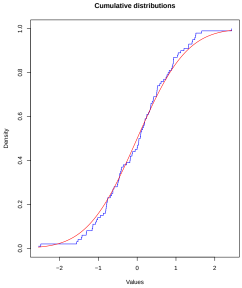
Running time of a software according to input size

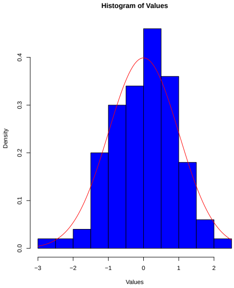
100 measures per size mean per size boxplot per size

Comparing two distributions : overlay graphics

Some methods to check if two distributions are close :

- Overlay cumulative distribution functions on the same graph
- Overlay histograms for well-chosen intervals
- Draw a Q-Q plot

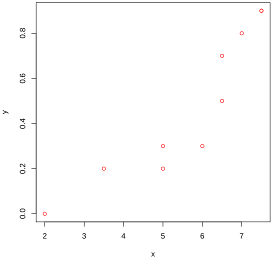




Overlaying an empirical distribution and a normal distribution

Comparing two distributions : Q-Q plot

Plot points (α-quantile 1st distrib,α-quantile 2nd distrib) for a set of well-chosen α (e.g., $\alpha = \frac{k}{n+1}$ for $1 \leq k \leq n$).

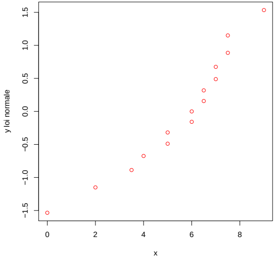


Example : two empirical distribution and $\alpha = \frac{1}{11}, \ldots, \frac{10}{11}$

Sample X	0.0	2.0	3.5	4.0	5.0	5.0	6.0	6.0	6.5	6.5	7.0	7.0	7.5	7.5	9.0
Sample Y		0.0	0.2		0.2	0.3		0.3	0.5	0.7		0.8	0.9	0.9	

Comparing two distributions : Q-Q plot

Plot points (α-quantile 1st distrib,α-quantile 2nd distrib) for a set of well-chosen α (e.g., $\alpha = \frac{k}{n+1}$ for $1 \leq k \leq n$).



Example : empirical distrib vs normal law $\mathcal{N}(0,1)$ et $\alpha = \frac{1}{n+1}, \ldots, \frac{n}{n+1}$

Sample X	0.0	2.0	3.5	4.0	5.0	5.0	6.0	6.0	6.5	6.5	7.0	7.0	7.5	7.5	9.0
Normal law Y	$q_{\frac{1}{16}}$	$q_{\frac{2}{16}}$	$q_{\frac{3}{16}}$	$q_{\frac{4}{16}}$	$q_{\frac{5}{16}}$	$q_{\frac{6}{16}}$	$q_{\frac{7}{16}}$	$q_{\frac{8}{16}}$	$q_{\frac{9}{16}}$	$q_{\frac{10}{16}}$	$q_{\frac{11}{16}}$	$q_{\frac{12}{16}}$	$q_{\frac{13}{16}}$	$q_{\frac{14}{16}}$	$q_{\frac{15}{16}}$

Inferential statistics : ingredients

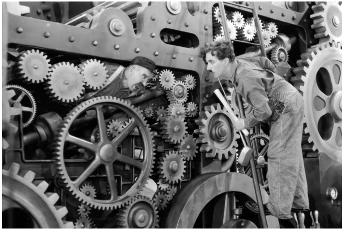
Data : a sample $(x_1, \ldots, x_n) \in E^n$
Models :

- parametric : chosen in a family of laws parametrized by one or several values θ
- non parametric : no restriction about the available laws

Question : assuming that data is driven/generated by one of the models considered, find the model(s) which best fit(s) the data ("best" yet to define)

Textbook case : a faulty machine

Scenario : a machine producing some devices sometimes functional (0), sometimes faulty (1).



Experiment : collecting a sample of $n = 100$ devices

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Textbook case : a faulty machine

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Model chosen : sample generated by an i.i.d. sequence of random variables $X_1, \dots, X_n$ with Bernoulli law of paramer p (unknown).

Question : can you give the exact value of p ? a range of values ? with some guarantees ? can you decide whether $p > p_0$ threshold from which production must be stopped ?

Textbook case : suggestions for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Textbook case : suggestions for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea 1 : $p = \frac{n_1}{n} = \frac{20}{100}$ where $n_1 =$ nb of 1 (strong law of large nb)

Textbook case : suggestions for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea 1 : $p = \frac{n_1}{n} = \frac{20}{100}$ where $n_1 =$ nb of 1 (strong law of large nb)

Idea 2 : proba of occurence of this sample $= \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1}$
→ choose p to maximize this proba : $p = \frac{n_1}{n} = \frac{20}{100}$

Textbook case : suggestions for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea 1 : $p = \frac{n_1}{n} = \frac{20}{100}$ where $n_1 =$ nb of 1 (strong law of large nb)

Idea 2 : proba of occurence of this sample $= \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1}$
$\rightarrow$ choose p to maximize this proba : $p = \frac{n_1}{n} = \frac{20}{100}$

Do we bet ?

Textbook case : suggestions for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea 1 : $p = \frac{n_1}{n} = \frac{20}{100}$ where $n_1 =$ nb of 1 (strong law of large nb)

Idea 2 : proba of occurence of this sample $= \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1}$
$\rightarrow$ choose p to maximize this proba : $p = \frac{n_1}{n} = \frac{20}{100}$

Do we bet ? Dangerous because no guarantee : for any $p \neq 0, \neq 1$, proba of occurence of this sample > 0.

Textbook case : a range with guarantees for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea : find some functions/algorithms I^- and I^+ from $\mathbb{R}^n$ tp $\mathbb{R}$ such that you can evaluate/bound $\mathbb{P}(p \in [I^-(X_1,\ldots,X_n), I^+(X_1,\ldots,X_n)])$ in an interesting way. If $\mathbb{P}(p \in [I^-(X_1,\ldots,X_n), I^+(X_1,\ldots,X_n)]) \geq \alpha$, the range is called *confidence interval* of *level* α.

Textbook case : a range with guarantees for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea 1 : Chebychev Inequality $\mathbb{P}(|X - \mathbb{E}(X)| \geq \delta) \leq Var(X)/\delta^2$
Here $\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - p| \geq \delta) \leq \frac{p(1-p)}{\delta^2}$

Textbook case : a range with guarantees for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea 1 : Chebychev Inequality $\mathbb{P}(|X - \mathbb{E}(X)| \geq \delta) \leq Var(X)/\delta^2$
Here $\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - p| \geq \delta) \leq \frac{p(1-p)}{\delta^2} \quad \geq \frac{1}{4n\delta^2}$
Thus $\mathbb{P}(p \in [\widehat{p_n} - \delta, \widehat{p_n} + \delta]) \geq 1 - \frac{1}{4n\delta^2}$ with $\widehat{p_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Choose δ such that $1 - \frac{1}{4n\delta^2} = \alpha$, that is $\delta = \frac{1}{2\sqrt{(1-\alpha)n}}$
Application : here to get a valid interval with proba $\alpha = 90\%$, use $\mathbb{P}(p \in [\widehat{p_{100}} - \frac{1}{\sqrt{40}}, \widehat{p_{100}} + \frac{1}{\sqrt{40}}]) = 0.9$, our sample interval $\approx$ $[0.04, 0.36]$

Textbook case : a range with guarantees for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea 2 : Central Limit Theorem

$\mathbb{P}(|\frac{\sqrt{n}}{\sqrt{Var(X)}}(\overline{X_n} - \mathbb{E}(X))| \leq \delta) \to \frac{1}{2\pi} \int_{-\delta}^{+\delta} e^{-x^2/2} dx$

Here $\mathbb{P}(|\frac{\sqrt{n}}{\sqrt{p(1-p)}}(\widehat{p_n} - p)| \leq \delta) \leq \mathbb{P}(|\widehat{p_n} - p| \leq \frac{\delta}{2\sqrt{n}})$

Let $\alpha = 0.9$, choose δ such that $\frac{1}{2\pi} \int_{-\delta}^{+\delta} e^{-x^2/2} dx = \alpha$, i.e., $\delta \approx 1.64$

Asymptotically $\mathbb{P}(p \in [\widehat{p_n} - \frac{1.64}{2\sqrt{n}}, \widehat{p_n} - \frac{1.64}{2\sqrt{n}}]) \geq 0.9$

Textbook case : a range with guarantees for p ?

Experiment : sample of size $n = 100$

00010 00000 11000 01000 10001
00000 00000 01110 00000 10000
00000 00000 01011 00000 00101
10000 00000 11011 00000 00000

Idea 2 : Central Limit Theorem

$\mathbb{P}(|\frac{\sqrt{n}}{\sqrt{Var(X)}}(\overline{X_n} - \mathbb{E}(X))| \leq \delta) \rightarrow \frac{1}{2\pi} \int_{-\delta}^{+\delta} e^{-x^2/2} dx$

Here $\mathbb{P}(|\frac{\sqrt{n}}{\sqrt{p(1-p)}}(\widehat{p_n} - p)| \leq \delta) \leq \mathbb{P}(|\widehat{p_n} - p| \leq \frac{\delta}{2\sqrt{n}})$

Let $\alpha = 0.9$, choose δ such that $\frac{1}{2\pi} \int_{-\delta}^{+\delta} e^{-x^2/2} dx = \alpha$, i.e., $\delta \approx 1.64$

Asymptotically $\mathbb{P}(p \in [\widehat{p_n} - \frac{1.64}{2\sqrt{n}}, \widehat{p_n} - \frac{1.64}{2\sqrt{n}}]) \geq 0.9$