Introduction to Probabilities and Statistics

Arnaud Legrand and Jean-Marc Vincent

Scientific Methodology and Performance Evaluation
ENS Lyon, November 2016

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

  A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement (or an observation)

  You can think of it as a small box:

  - Every time you open the box, you get a different value.
  - I will use this box analogy throughout the whole lecture and I encourage you to ask yourself what the box can be in your own studies

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

  A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement (or an observation)

  You can think of it as a small box:

  - Every time you open the box, you get a <u>different</u> value.
  - I will use this box analogy throughout the whole lecture and I encourage you to ask yourself what the box can be in your own studies
- Formally a probability space is defined by $(\Omega, \mathcal{F}, \mathrm{P})$ where:
  - Ω, the sample space, is the set of all possible outcomes
    - E.g., all the possible combinations of your DNA with the one of your {girl|boy}friend
    - You may or may not be able to observe directly the outcome.

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

  A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement (or an observation)

  You can think of it as a small box:

  - Every time you open the box, you get a different value.
  - I will use this box analogy throughout the whole lecture and I encourage you to ask yourself what the box can be in your own studies

- Formally a probability space is defined by $(\Omega, \mathcal{F}, \mathrm{P})$ where:
  - Ω, the sample space, is the set of all possible outcomes
    - E.g., all the possible combinations of your DNA with the one of your {girl|boy}friend
    - You may or may not be able to observe directly the outcome.
  - $\mathcal{F}$ if the set of events where an event is a set containing zero or more outcomes
    - E.g., the event of "the DNA corresponds to a girl with blue eyes"
    - An event is somehow more tangible and can generally be observed

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

  A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement (or an observation)

  You can think of it as a small box:
  - Every time you open the box, you get a different value.
  - I will use this box analogy throughout the whole lecture and I encourage you to ask yourself what the box can be in your own studies
- Formally a probability space is defined by $(\Omega, \mathcal{F}, \mathrm{P})$ where:
  - Ω, the sample space, is the set of all possible outcomes
    - E.g., all the possible combinations of your DNA with the one of your {girl|boy}friend
    - You may or may not be able to observe directly the outcome.
  - $\mathcal{F}$ if the set of events where an event is a set containing zero or more outcomes
    - E.g., the event of "the DNA corresponds to a girl with blue eyes"
    - An event is somehow more tangible and can generally be observed
  - The probability measure $\mathrm{P} : \mathcal{F} \to [0, 1]$ is a function returning an event's probability (P("having a brown-eyed baby girl") = 0.0005)

Continuous random variable

- A random variable associates a numerical value to outcomes

$$X : \Omega \to \mathbb{R}$$

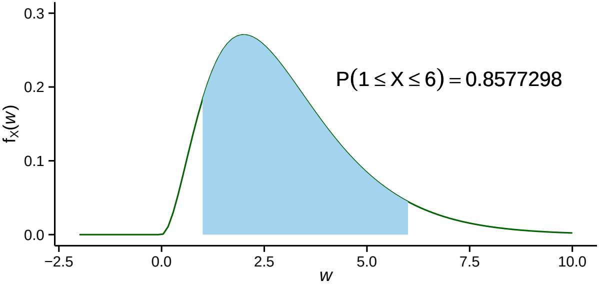
  - E.g., the weight of the baby at birth (assuming it solely depends on DNA, which is quite false but it's for the sake of the example)
  - Since many computer science experiments are based on time measurements, we focus on continuous variables
- **Note:** To distinguish random variables, which are complex objects, from other mathematical objects, they will always be written in blue capital letters in this set of slides (e.g., X)
- The probability measure on Ω induces probabilities on the values of X
  - $\mathrm{P}(X = 0.5213)$ is generally 0 as the outcome never exactly matches
  - $\mathrm{P}(0.5213 \leq X \leq 0.5214)$ may however be non-zero

Probability distribution

A probability distribution (a.k.a. probability density function or p.d.f.) is used to describe the probabilities of different values occurring

- A random variable X has density f_X, where f_X is a non-negative and integrable function, if:

$$\boxed{P[a \leq X \leq b] = \int_a^b f_X(w)\, dw}$$



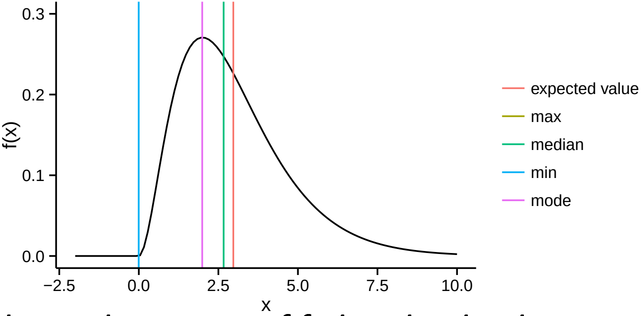
Note: the X in $1 \leq X \leq 6$ *should be in blue...*

- **Note**: people often confuse the sample space with the random variable. Try to make the difference when modeling your system, it will help you

Characterizing a random variable

The probability density function fully characterizes the random variable but it is also complex object

- It may be symmetrical or not
- It may have one or several modes
- It may have a bounded support or not, hence the random variable may have a minimal and/or a maximal value
- The median cuts the probabilities in half



These are interesting aspects of f_X but they barely summarize it

Expected value and variance

- When one speaks of the "expected price", "expected height", etc. one means the expected value of a random variable that is a price, a height, etc.

  $$\mathrm{E}[X] = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k = \int_{-\infty}^{\infty} x f_X(x)\ \mathrm{d}x$$

  The expected value of X is the “average value” of X.

  It is **not** the most probable value. The mean is <u>one</u> aspect of the distribution of X. The median or the mode are other interesting aspects.
- The variance is a measure of how far the values of a random variable are spread out from each other.

  If a random variable X has the expected value (mean) $\mu = \mathrm{E}[X]$, then the variance of X is given by:

  $$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\ \mathrm{d}x$$
- The standard deviation σ is the square root of the variance. This normalization allows to compare it with the expected value

Outline

1. Brief reminder on probabilities

2. Moment Generating Function
   - Intuitions
   - Properties

3. Toward the Central Limit Theorem
   - Law of Large Numbers
   - Central Limit Theorem
   - Central Limit Theorem consequences

Definition

Working with the density function is not always convenient, especially when summing random variables (it implies convolving the pdf). We need an *alternate representation*.

How could we summarize a random variable ?

- By its mean, its variance, its skewness, . . . by its moments $\mu_k = \mathrm{E}(X^k)$
- It is not clear that it would be sufficient although we would know a lot about f_X.

Let's define the moment generating function $\mathrm{M}_X(t)$ as follows:

$$\mathrm{M}_X(t) = \mathrm{E}\left(e^{tX}\right) = \mathrm{E}\left(\sum_{k=0}^{\infty} \frac{t^k X^k}{k!}\right) = \mathrm{E}\left(\sum_{k=0}^{\infty} \frac{t^k X^k}{k!}\right) = \sum_{k=0}^{\infty} \mu_k \frac{t^k}{k!}$$

$$= \int e^{tx} f_X(x) dx$$

Deriving moments with the mgf

Remember we have $\mathrm{M}_{X}(t) = \sum_{k=0}^{\infty} \mu_k \frac{t^k}{k!}$

Therefore $\frac{d^n \, \mathrm{M}_{X}}{dt^n}(0) = \mu_n$

All the moments of X are encoded in $\mathrm{M}_{X}(t)$. Is there more ?

Characterization of a distribution through the mgf

Let's assume that X is discrete $\big((x_1, p_1), \dots, (x_n, p_n)\big)$ with $x_1 < \cdots < x_n$

- Then $\mathrm{M}_X(t) = \mathrm{E}\left(e^{tX}\right) = \sum_{j=1}^n p_j e^{tx_j} = \sum_{j=1}^n p_j (e^t)^{x_j}$
- Therefore $\mathrm{M}_X(t) \underset{t\to\infty}{\sim} p_n e^{tx_n}$ and $\mathrm{M}'_X(t) \underset{t\to\infty}{\sim} p_n x_n e^{tx_n}$.
  $\rightsquigarrow \frac{\mathrm{M}'_X(t)}{\mathrm{M}_X(t)} \xrightarrow[t\to\infty]{} x_n$
- Hence, we can determine x_n, then p_n, substract $p_n e^{tx_n}$ from $\mathrm{M}_X(t)$ and proceed to find x_{n-1}.

X is fully characterized by its mgf M_X

Proving the same results when X is continuous, requires to go through Fourier transform.

Outline

1. Brief reminder on probabilities

2. Moment Generating Function
   - Intuitions
   - Properties

3. Toward the Central Limit Theorem
   - Law of Large Numbers
   - Central Limit Theorem
   - Central Limit Theorem consequences

Convenient properties

$$\begin{aligned} \mathrm{M}_{aX+b}(t) &= \mathrm{E}\left(e^{t(aX+b)}\right) = \mathrm{E}\left(e^{bt}e^{atX}\right) \\ &= e^{bt}\,\mathrm{M}_X(at) \end{aligned}$$

$$\begin{aligned} \mathrm{M}_{X+Y}(t) &= \mathrm{E}\left(e^{t(X+Y)}\right) = \mathrm{E}\left(e^{tX+tY}\right) = \mathrm{E}\left(e^{tX}e^{tY}\right) = \mathrm{E}\left(e^{tX}\right)\mathrm{E}\left(e^{tY}\right) \\ &= \mathrm{M}_X(t).\,\mathrm{M}_Y(t) \end{aligned}$$

Mgf of usual laws

- Uniform law: $\mathrm{M}_X(t) = \begin{cases} \frac{e^{tb}-e^{ta}}{t(b-a)} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases}$
- Exponential law: $f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$

$$\mathrm{M}_X(t) = \mathrm{E}\left(e^{tX}\right) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx = \lambda \int_0^\infty e^{(t-\lambda)x}dx$$

$$= \lambda \left[\frac{e^{(t-\lambda)x}}{t-\lambda}\right]_0^\infty = \frac{\lambda}{\lambda - t} \qquad (\text{for } t < \lambda)$$

This allows to easily compute moments and sum random variables.
The moment generating function is somehow similar to the Fourier transform on periodic signals.

Outline

1 Brief reminder on probabilities

2 Moment Generating Function
- Intuitions
- Properties

3 Toward the Central Limit Theorem
- Law of Large Numbers
- Central Limit Theorem
- Central Limit Theorem consequences

How to estimate the Expected value?

To empirically estimate the expected value of a random variable X, one repeatedly measures observations of the variable and computes the arithmetic mean of the results

This is called the sample mean and it intuitively converges to the expected value

Unfortunately, if you repeat the estimation, you may get a different value since X is a random variable ...

What can we really say ?

On the way to the Law of Large Numbers

Chebyshev Inequality

Let X be a random variable with expected value $\mu = \mathrm{E}(X)$, and let $\varepsilon > 0$ be any positive real number. Then $\mathrm{P}(|X - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}$.

Proof

$$\begin{aligned}\mathrm{Var}(X) = \int (x-\mu)^2 f(x).dx &\geq \int_{|x-\mu|\geq\varepsilon} (x-\mu)^2 f(x).dx \\ &\geq \int_{|x-\mu|\geq\varepsilon} \varepsilon^2 f(x).dx = \varepsilon^2 \underbrace{\int_{|x-\mu|\geq\varepsilon} f(x).dx}_{\mathrm{P}(|X-\mu|\geq\varepsilon)}\end{aligned}$$

Law of Large Numbers

Law of Large Numbers

Let X_1, X_2, ..., X_n be a sequence of identical and independent random variables with finite expected value $\mu = \mathrm{E}(X_i)$ and finite variance $\sigma^2 = \mathrm{Var}(X_i)$. Let $S_n = X_1 + X_2 + \cdots + X_n$.
Then for any $\varepsilon > 0$, $\mathrm{P}(|S_n/n - \mu| \geq \varepsilon) \xrightarrow[n\to\infty]{} 0$.

Proof

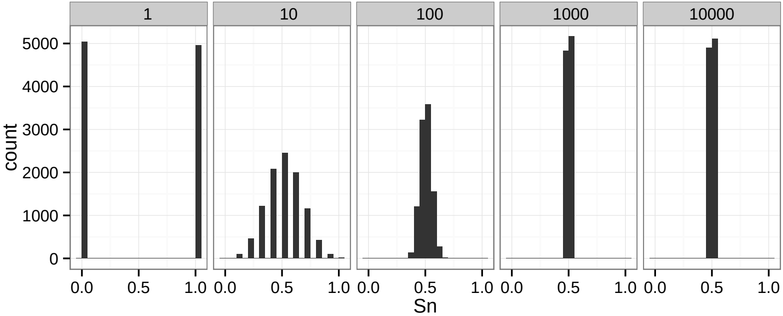
The X_i are i.i.d, hence:

- $\mathrm{Var}(S_n) = n.\sigma^2 \rightsquigarrow \mathrm{Var}(S_n/n) = \sigma^2/n$.
- $\mathrm{E}(S_n/n) = \mu$.

Using Chebyshev's inequality:

$$\mathrm{P}(|S_n/n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0 \text{ (for a fixed } \varepsilon\text{)}$$

Illustration: convergence in probability



So we do converge to a spike, but how ?
Assume $\sigma = 1$ and we aim at having a precision of $\varepsilon = .1$. For $n = 500$, the previous formula only gives us
$P(|S_n/n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} = \frac{100}{n} = 0.5$ ☹

Outline

1. Brief reminder on probabilities

2. Moment Generating Function
   - Intuitions
   - Properties

3. Toward the Central Limit Theorem
   - Law of Large Numbers
   - Central Limit Theorem
   - Central Limit Theorem consequences

Central Limit Theorem [CLT]

- Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size n (i.e., a sequence of independent and identically distributed random variables with expected values μ and variances σ^2)
- We know that $E(S_n/n) = \mu$ and Var (S_n) = n$\sigma^2.
- Let's define the standardized mean of these random variables as:

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

  We have $E(S_n^*) = 0$ and $\mathrm{Var}(S_n^*) = 1$.
- For large n's, the distribution of S_n^* is approximately normal

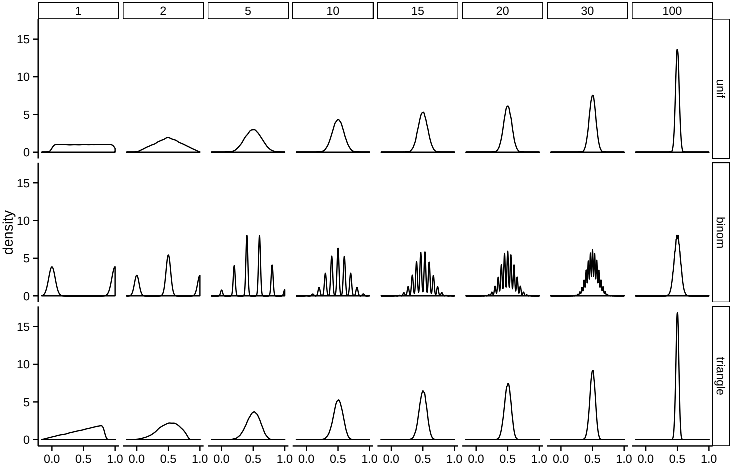
$$S_n^* \xrightarrow[n\to\infty]{} \mathcal{N}(0,1)$$

  Or equivalently

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

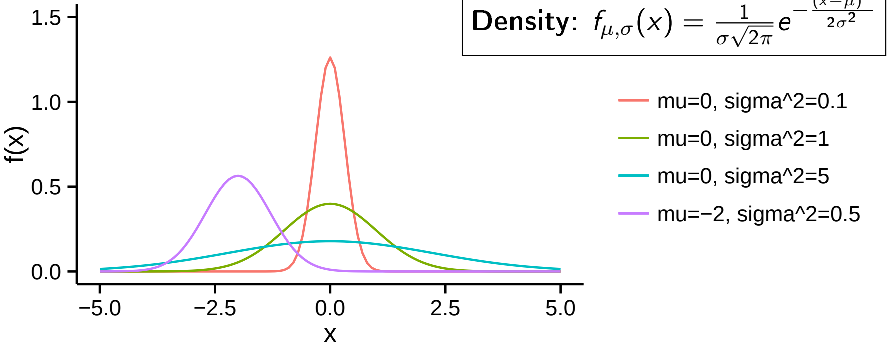
CLT Illustration: the mean smooths distributions

Start with an arbitrary distribution and compute the distribution of S_n for increasing values of n.



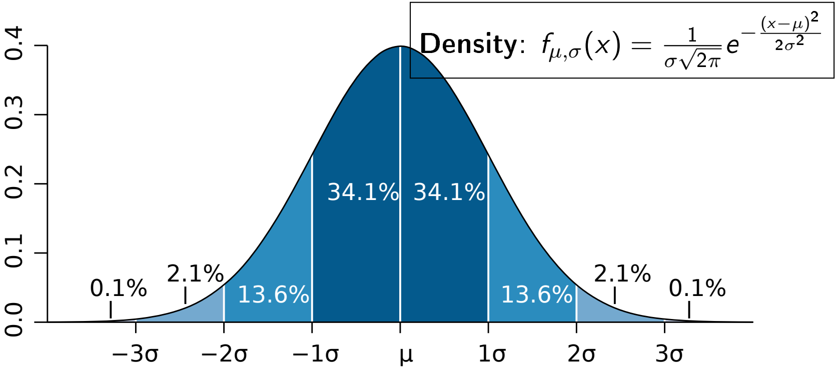
The Normal distribution

Density: $f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



The smaller the variance the more “spiky” the distribution.

The Normal distribution



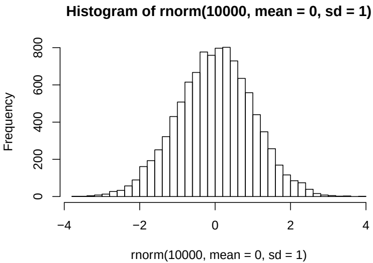
Density: $f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

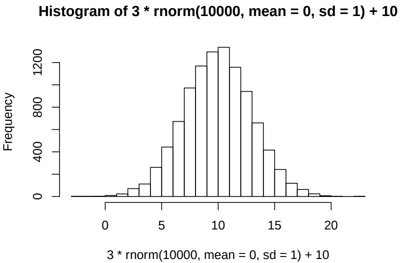
The smaller the variance the more "spiky" the distribution.

- Dark blue is less than one standard deviation from the mean$\approx$ 68% of the set.
- Two standard deviations from the mean (medium and dark blue)$\approx$95%
- Three standard deviations (light, medium, and dark blue)$\approx$99.7%

The Normal distribution (property 1)

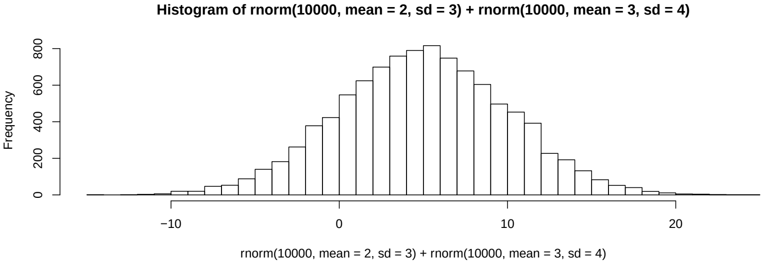
The family of normal distributions is closed under linear transformations: if X is normally distributed with mean μ and standard deviation σ, then the variable $Y = aX + b$ is also normally distributed, with mean $a\mu + b$ and standard deviation $|a|\sigma$.


Histogram of rnorm(10000, mean = 0, sd = 1)


Histogram of 3 * rnorm(10000, mean = 0, sd = 1) + 10

The Normal distribution (property 2)

Convolution: if X_1 and X_2 are two independent normal random variables, with means μ_1, μ_2 and standard deviations σ_1, σ_2, then their sum $X_1 + X_2$ will also be normally distributed, with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.



Intuitively, if S_n^* converges to something (say $\mathcal{L}$), it "*has to*" be a normal distribution:

$$\frac{1}{2}(\underbrace{S_{1\ldots n}^*}_{\sim\mathcal{L}} + \underbrace{S_{n+1\ldots 2n}^*}_{\sim\mathcal{L}}) = \underbrace{S_{2n}^*}_{\sim\mathcal{L}}$$

Moment generating function of the normal distribution

Let's assume $X \sim \mathcal{N}(0,1)$.

$$\begin{aligned}
\mathrm{M}_X(t) &= \int e^{tx} f_{\mathcal{N}}(x).dx = \int e^{tx} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \int \frac{e^{\frac{1}{2}(-x^2+tx)}}{\sqrt{2\pi}} dx \\
&= \int \frac{e^{\frac{1}{2}(-(x-t)^2+t^2)}}{\sqrt{2\pi}} dx = e^{\frac{x^2}{2}} \int \frac{e^{\frac{-(x-t)^2}{2}}}{\sqrt{2\pi}} dx = e^{\frac{x^2}{2}} \int \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}} dx \\
&= e^{\frac{x^2}{2}}
\end{aligned}$$

Actually, if we assume $X \sim \mathcal{N}(\mu, \sigma^2)$, one can easily prove in the same way that:

$$\mathrm{M}_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Proof of the CLT

$$\boxed{\begin{array}{l} \mathrm{M}_X(t) = \mathrm{E}(e^{tX}) \approx 1 + \mu t + \sigma^2 \frac{t^2}{2} + o(t^2) \\ \rightsquigarrow \log(\mathrm{M}_{X-\mu}(t)) \approx \sigma^2 \frac{t^2}{2} + o(t^2) \end{array}}$$

$$\boxed{\begin{cases} S_n = X_1 + \cdots + X_n \\ S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}} \end{cases}}$$

We have:

$$\mathrm{M}_{S_n^*}(t) = \mathrm{E}(e^{tS_n^*}) = \mathrm{E}(e^{t\frac{S_n - n\mu}{\sqrt{n}\sigma}}) = \mathrm{E}(e^{\frac{t}{\sigma\sqrt{n}}(S_n - n\mu)}) = \mathrm{M}_{S_n - n\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)$$

$$= \left(\mathrm{M}_{X-\mu}\left(\underbrace{\frac{t}{\sigma\sqrt{n}}}_{\xrightarrow[n\to\infty]{}0}\right)\right)^n \qquad \text{(since } \mathrm{M}_{X+Y}(t) = \mathrm{M}_X(t)\,\mathrm{M}_Y(t)\text{)}$$

$$= \exp\left(n\log\left(\mathrm{M}_{X-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)\right) = \exp\left(n\left(\sigma^2\frac{t^2}{2n\sigma^2} + o\left(\frac{t^2}{n^2}\right)\right)\right)$$

$$= \exp\left(\frac{t^2}{2} + o(t^2/n)\right) \xrightarrow[n\to\infty]{} e^{t^2/2}, \text{ which is the mgf of } \mathcal{N}(0,1) \qquad \square$$

CLT = convergence of laws

The law of S_n^* converges to $\mathcal{N}(0,1)$. In other words, whatever the initial law of X:

$$\lim_{n\to\infty} \mathrm{P}[a < S_n^* < b] = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2} dx$$

It provides a reasonable approximation when close to the peak of the normal distribution.
(it requires a very large number of observations to stretch into the tails)

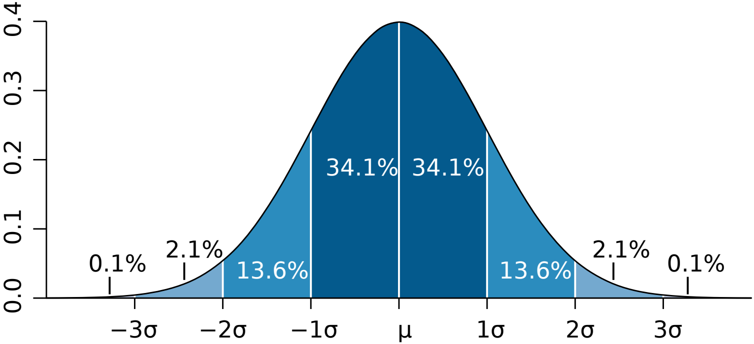
Outline

1. Brief reminder on probabilities

2. Moment Generating Function
   - Intuitions
   - Properties

3. Toward the Central Limit Theorem
   - Law of Large Numbers
   - Central Limit Theorem
   - Central Limit Theorem consequences

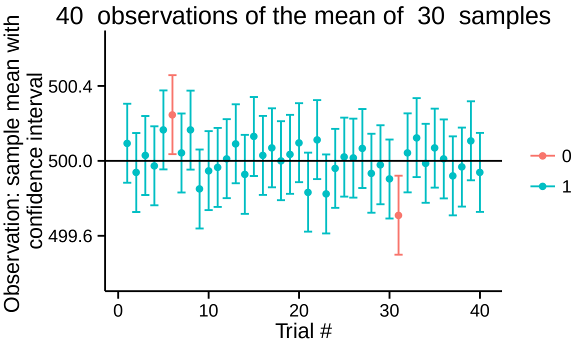
Confidence interval



When n is large:

$$P\left(\mu \in \left[S_n - 2\frac{\sigma}{\sqrt{n}}, S_n + 2\frac{\sigma}{\sqrt{n}}\right]\right) = P\left(S_n \in \left[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right]\right) \approx 95\%$$

Confidence interval



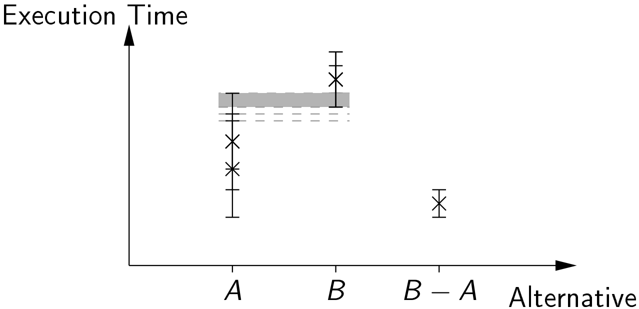
When n is large:

$$\mathrm{P}\left(\mu \in \left[S_n - 2\frac{\sigma}{\sqrt{n}}, S_n + 2\frac{\sigma}{\sqrt{n}}\right]\right) = \mathrm{P}\left(S_n \in \left[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right]\right) \approx 95\%$$

There is 95% of chance that the true mean lies within $2\frac{\sigma}{\sqrt{n}}$ of the sample mean.

Without any particular hypothesis

- Assume, you have evaluated two alternatives A and B on n different setups
- You therefore consider the associated random variables A and B and try to estimate their expected values μ_A and μ_B

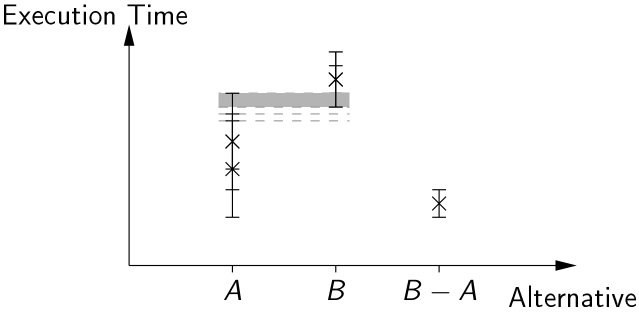


The two 95% confidence intervals do not overlap

$\rightsquigarrow \mu_A < \mu_B$ with more than 90% of confidence ☺

Without any particular hypothesis

- Assume, you have evaluated two alternatives A and B on n different setups
- You therefore consider the associated random variables A and B and try to estimate their expected values μ_A and μ_B



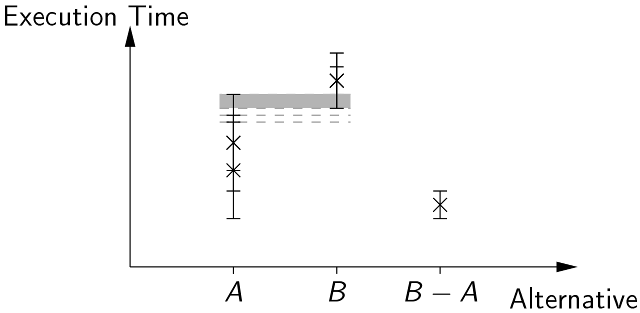
The two 95% confidence intervals do overlap

⇝ Nothing can be concluded ☹

Reduce C.I?

Without any particular hypothesis

- Assume, you have evaluated two alternatives A and B on n different setups
- You therefore consider the associated random variables A and B and try to estimate their expected values μ_A and μ_B

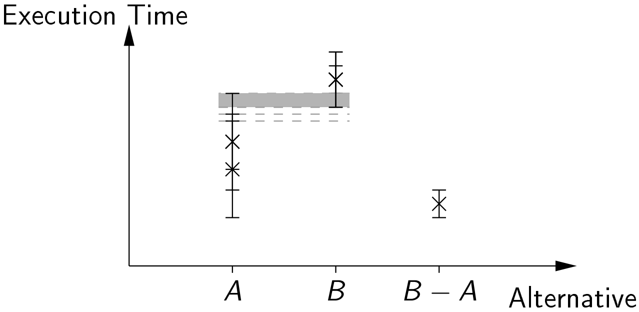


The two 70% confidence intervals do not overlap

$\rightsquigarrow$ $\mu_A < \mu_B$ with less than 50% of confidence ☹ $\rightsquigarrow$ more experiments...

Without any particular hypothesis

- Assume, you have evaluated two alternatives A and B on n different setups
- You therefore consider the associated random variables A and B and try to estimate their expected values μ_A and μ_B



The width of the confidence interval is proportional to $\frac{\sigma}{\sqrt{n}}$

You can estimate how much more experiments you need☺

4 times more to halve it! ☹ Try to reduce variance if you can...☺