Introduction to Probabilities and Statistics

Arnaud Legrand and Jean-Marc Vincent

Scientific Methodology and Performance Evaluation
M2R MOSIG, Grenoble, September-December 2015

Outline

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 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.

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.
  - $\mathcal{F}$ 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 <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.
  - $\mathcal{F}$ 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 $\mathrm{P} : \mathcal{F} \rightarrow [0, 1]$ 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

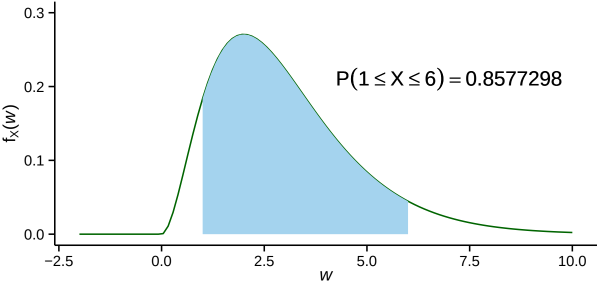
$$X : \Omega \to \mathbb{R}$$

  - 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
  - $P(X = 0.5213)$ is generally 0 as the outcome never exactly matches
  - $P(0.5213 \leq X \leq 0.5214)$ 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: $\boxed{P[a \leq X \leq b] = \int_a^b f_X(w)\, dw}$



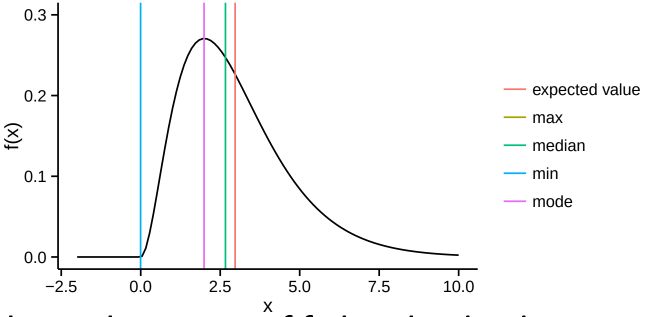
Note: the X in $1 \leq X \leq 6$ *should be in blue...*

- **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

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

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

Central Limit Theorem [CLT]

- Let $\{X_1, X_2, \dots, X_n\}$ 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)
- The sample mean of these random variables is:

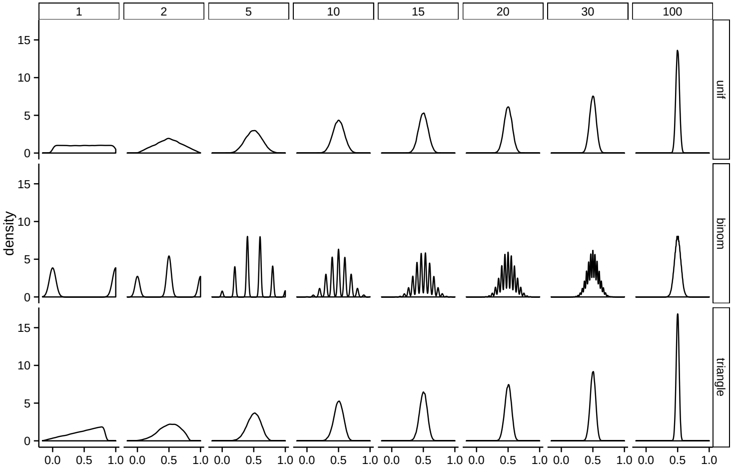
$$S_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

  S_n is a random variable too!
- It is unbiased, i.e., $\mathrm{E}[S_n] = \mathrm{E}[X]$
- For large n's, the distribution of S_n is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$

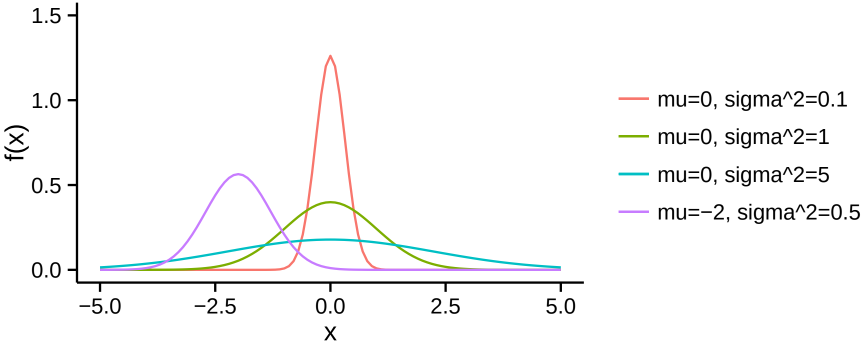
$$S_n \xrightarrow[n \to \infty]{} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

CLT Illustration: the mean smooths distributions

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

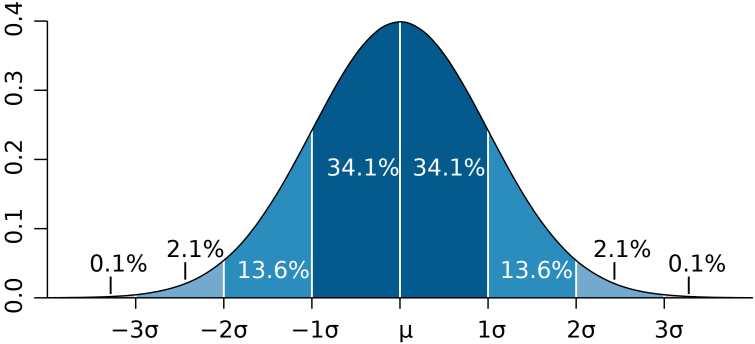


The Normal Distribution



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

The Normal Distribution

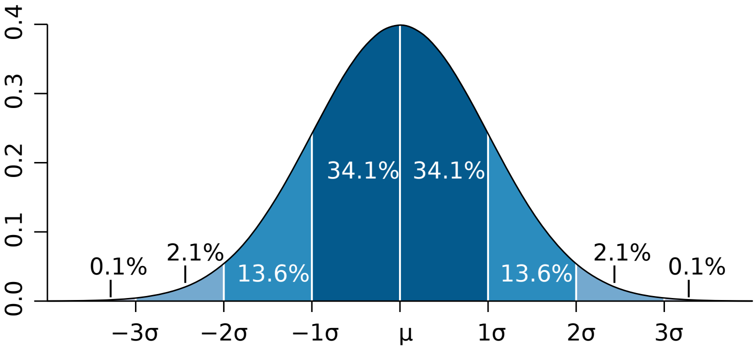


The smaller the variance the more "spiky" the distribution.

- Dark blue is less than one standard deviation from the mean. For the normal distribution, this accounts for about 68% of the set.
- Two standard deviations from the mean (medium and dark blue) account for about 95%
- Three standard deviations (light, medium, and dark blue) account for about 99.7%

Outline

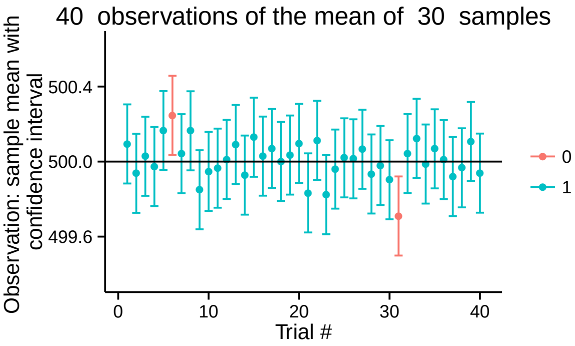
CLT consequence: confidence interval



When n is large:

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

CLT consequence: confidence interval



When n is large:

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

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