Probability, Statistics, And Population Genetics: A Brief Introduction

Ziyang Xia

April 9, 2021

Background: modelling randomness in science

- Classical approach of natural science - modelling NATURE without randomness
- "God does not play dice." — Albert Einstein
- $x \rightarrow y$: a causal path
- $y = f(x, \theta)$: a mathematical model
  - x: independent variable - *causes*
  - θ: parameter - inherent characteristics of the model
  - y: dependent variable - *effects*
- Two problems of such an approach:
  - Inherent randomness of studied objects
    e.g., quantum mechanisms
  - Fail of reductionism in complex systems - too much potential causes
    e.g., biological organism or society
- Pervasive randomness in molecular (e.g., mutation, recombination) and behavioral (e.g., mating, migration) processes of population genetics
- Modelling the randomness - pivotal issue for population genetics.

To quantify the randomness: random variable & probability

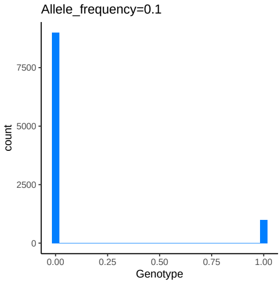
- $X = x_1, x_2, x_3, \ldots$
- X **random variable**: a quantity with different possible results
- $X = x_i$ **random event**: each feasible value X can be
  - Tossing a coin: $X = \{0, 1\}$
  - Observing genotype of a haployid individual: $X = \{0, 1\}$
- Ω **universe**: set contains all the possible x_i for X
- In statistics, actually observed $X = x_i$ are **data**
- Each random event has some "degree of plausibility" to happen
- This can be quantified by a value called **probability** $Pr(x_i)$
  - Tossing a coin: $Pr(X = 0) = Pr(X = 1) = \frac{1}{2}$ (usually but not always)
  - Observing genotype of a haployid individual: $Pr(X = 0) = 1 - p$, $Pr(X = 1) = p$, where p is derived allele frequency of the population

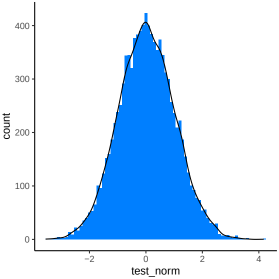
Function of random variable & probability I

- We can use function to describe the correspondance between random events and probabilities
- $f(X, \theta)$, $\forall x_i \in \Omega$, $f(X = x_i, \theta) = Pr(x_i)$
- **probability distribution**: the function with random variable X as **independent**, probability $Pr(X)$ as **dependent**
- i.e., Mapping of each random event to corresponding value of probability: $f(X, \theta) : X \to Pr(X)$
- θ is still **parameter**, which defines characteristics of probability distribution.

Function of random variable & probability II

- For both examples, $f(X, \theta) = p^X(1-p)^{1-X}$, where $X = \{0, 1\}$, $\theta = \{p\}$
- Note: θ is the set of all the parameters - there can be multiple parameters, e.g., mean & variance.
- Probability distribution is also known as **probability density function** (pdf) when the number of random events is infinite
  - Then the probability is integration: $Pr(a \leq X \leq b) = \int_a^b f(x, \theta)dx$



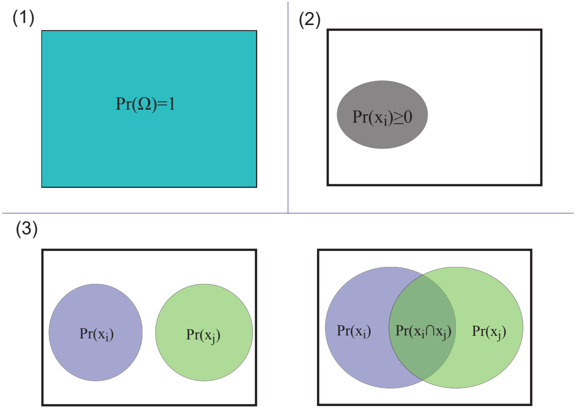


Kolmongorov's axioms of probability: formal definition I

- Probability can be formally defined by the following axioms:
  - (1) For $\forall x_i \in \Omega$, $Pr(x_i) \geqslant 0$
  - (2) $Pr(\Omega) = 1$
  - (3) For $\forall x_i, x_j \in \Omega$, $Pr(x_i \cup x_j) = Pr(x_i) + Pr(x_j)$, if $x_i \cap x_j = \varnothing$
    $$\iff Pr(x_i \cup x_j) = Pr(x_i) + Pr(x_j) - Pr(x_i \cap x_j)$$
- Probability theory is a self-consistent, pure mathematical system in ideal world
- Accordant with our logic for "true/false", "existance/inexistance", or "occurance/non-occurance" in both ideal and real worlds:
  - (1) & (2) $\iff$ 0: false; 1: true; $(0, 1)$: intermediate state = degree of plausibility
  - (3) $\iff$ $x \vee y = x + y - x \wedge y$ in logic/boolean algebra
  - Probability: extenstion of logic algebra from $\{0, 1\}$ to $[0, 1]$.

Kolmongorov's axioms of probability: formal definition II

- Since probability is an extention of logic algebra, we can just use the relationship between sets (of random events, e.g., x_i, x_j) to represent their probabilities (e.g., $Pr(x_i)$, $Pr(x_j)$)
  - i.e., relationship of sets $\iff$ relationship of probabilities
  - Probability: mass point ($X = x_i$); area ($a \leq X \leq b$)

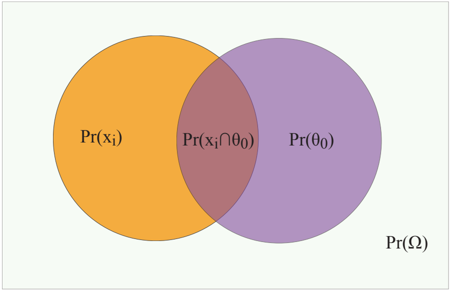


Relationship: probability theory & statistics I

- **Probability theory** only needs Kolmongorov's axioms as pre-assumptions - not bothered by real world.
  - In probability theory, for tossing a coin, p can be anything $\in [0,1]$, not necessarily $\frac{1}{2}$
- **Statistics** takes probability theory as its mathematical foundation, but also needs actually observed **data** to explain & predict the real world by **estimation**
  - In statistics, for tossing a coin, p is usually estimated to be around $\frac{1}{2}$ by observing sufficient data
- Different feasible types of probabilities conforming Kolmongorov's axioms
- $\implies$ Different corresponding types of statistics explaining the real world, e.g., classical/frequentist v.s. Bayesian

Conditional probaility

- Different choices of domain of definition in functions $\Longrightarrow$ different choices of Ω for probability definition
- **Conditional probability** $Pr(x_i \mid \theta_0)$: probability of $x = x_i$ when the event $\theta = \theta_0$ occurred or is assumed to occur
- $$Pr(x_i \mid \theta_0) = \frac{Pr(x_i \cap \theta_0)}{Pr(\theta_0)}, \qquad f(X \mid \theta) = \frac{f(X, \theta)}{f(\theta)}$$
- All the random events in Ω must meet the condition $\theta = \theta_0$; any event with $\theta \neq \theta_0 \notin \Omega \iff Pr(\theta_0 \mid \theta_0) = 1$



Independence

- If $Pr(x_i \mid \theta_0) = Pr(x_i) \iff Pr(x_i \cap \theta_0) = Pr(x_i)Pr(\theta_0)$, the occurrance of $\theta = \theta_0$ or not does not affect the probability of $x = x_i$
- In this case, we call $x = x_i$ and $\theta = \theta_0$ are mutually **independent** random events.
- An intuitive explanation: knowing the parameter $\theta = \theta_0$ does not provide further information of $x = x_i$
- Similarly, if $f(X \mid \theta) = f(X) \iff f(X, \theta) = f(X)f(\theta)$, then the two random variables X and θ are mutually independent

Bayes' theorem

- Given the definition of conditional probability, we have **Bayes' theorem**:
-
$$Pr(x_i \cap \theta_0) = Pr(x_i \mid \theta_0)Pr(\theta_0) = Pr(\theta_0 \mid x_i)Pr(x_i)$$

$$\iff Pr(\theta_0 \mid x_i) = \frac{Pr(\theta_0)Pr(x_i \mid \theta_0)}{Pr(x_i)}$$

- In the form of probability distribution:

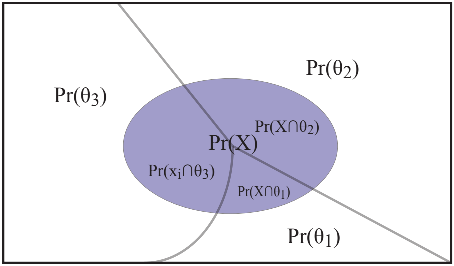
$$f(\theta \mid X) = \frac{f(\theta)f(X \mid \theta)}{Pr(X)}$$

- Bayes' theorem is the foundation of Bayesian statistics.
- But it is always true in probability theory & all kinds of statistics.

Total probability theorem

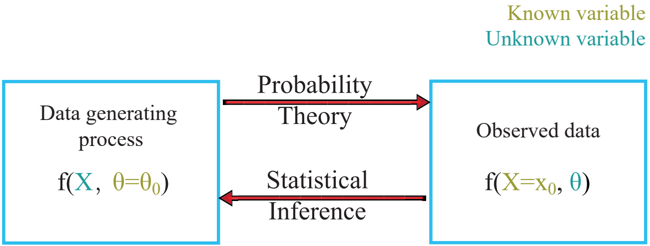
- **Total probability** of observing data X can be expressed as:
- $$Pr(X) = \sum_i Pr(X \cap \theta_i)$$
  where $\forall \theta_i \cap \theta_j = \varnothing$ and $\bigcup_i \theta_i = \Omega$
- Given conditional probability:

$$Pr(X) = \sum_i Pr(X \mid \theta_i) Pr(\theta_i), \qquad f(X) = \int f(X \mid \theta) f(\theta) d\theta$$



Relationship: probability theory & statistics II

- **Probability theory**
  - "Data generating machine": $f(X \mid \theta = \theta_0)$
  - $\theta \Rightarrow \{x_i\}$
  - Conditional on given parameter θ, generating a series of data x_i
- **Statistics**
  - $\{x_i\} \Rightarrow \theta$
  - Given known data x_i, estimating the value/range of parameter θ & original probability distribution
  - Parameter θ defines characteristic of a probability distribution



Statistical inference: Bayesian vs Frequentist

- **Frequentist**: parameter θ has an only real value
  - A stable data generating machine
  - As number of trial $n \to \infty$, frequency of observed data $X \to$ real probability defined by parameter θ
  - θ is a fixed value and does not have any probability distribution, so $f(\theta)$ or $f(\theta \mid X)$ is invalid for frequentists
- **Bayesian**: θ is a random variable with uncertainty
  - A data generating machine with instability (inherent or due to human observation)
  - θ should be model by probability distribution
  - **Prior** $f(\theta)$: θ's distribution without any further information
  - **Posterior** $f(\theta \mid X)$: θ's distribution with information from data X (i.e., knowing X occurred)
  - Using Bayes' theorem to gain posterior distribution - why "Bayesian" is named

Expectation: core parameter of probability distribution

- For a random variable X, its expectation (i.e., mean, average) is notated as μ or $\mathbb{E}[X]$
- Rationale: using a single parameter μ to describe all the possible random events x_i of random variable X, given their respective probability $f(x_i)$
- $$\sum_i f(x_i)x_i = \sum_i f(x_i)\mu \iff \mathbb{E}[X] = \mu = \sum_i f(x_i)x_i$$
- μ is the equivalent substitution of $\forall x_i$
- Extended definition of expectation: $\mathbb{E}[g(X)] = \sum_i f(x_i)g(x_i)$, where $g(X)$ is a function of X
- Particularly, $\mathbb{E}[c] = \sum_i f(x_i)c = c$ when c is a constant

Variance: dispersal from expectation

- Analogically, we can calculate the expectation of each x_i of random variable X away from mean μ, which is **variation**
- $$Var(X) = \sigma^2 = \mathbb{E}[(X-\mu)^2]$$
  $$= \mathbb{E}[X^2] - 2\mu\,\mathbb{E}(X) + \mu^2$$
  $$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$
- $\sigma = \sqrt{Var(X)}$: standard deviation

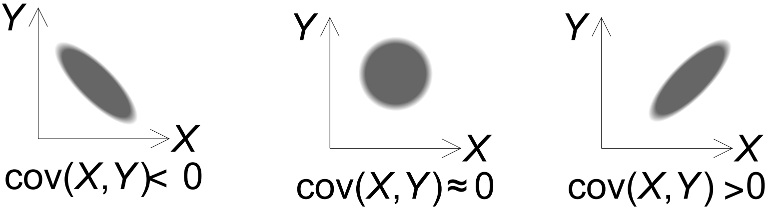
Covariance: variance of two variables I

- Variance: how a single random variable varies
- Analogical to variance, we define **covariance** as
- $$Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$
  $$= \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] - \mu_Y \mathbb{E}[X] + \mu_X \mu_Y$$
  $$= \mathbb{E}[XY] - E[X]E[Y]$$
- Particularly, $Cov(X, X) = Var(X)$
- Normalized covariance - Pearson's correlation coefficient:

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Covariance: variance of two variables II

- Covariance describes the varying tendency of Y when X varies:
  - $Cov(X, Y) > 0$: when X varies, Y tends to (i.e., has more probability to) vary in the same direction
  - $Cov(X, Y) < 0$: when X varies, Y tends to vary in the opposite direction
  - $Cov(X, Y) \approx 0$: when X varies, Y can vary in any direction - variation of X is nearly unrelated to variation of Y, X and Y are approximately independent



Independence of two variables

- Reviewing the definition of independence:
- Two random variables X, Y are independent $\implies f(X,Y) = f(X)f(Y)$ $\iff \mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ (Bertsekas, p.99)
- Therefore, if X and Y are independent,

$$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

- Note: Converse ($Cov(X,Y) = 0 \implies X$ and Y are independent) may not be true

Some important probability distributions I

- Bernoulli $X \sim Bern(p)$: a single trial with probability p
  - e.g., observing derived allele in a haploid individual
  - $f(X, \theta) = p^X(1-p)^{1-X}$
  - $\mathbb{E}(X) = p, Var(X) = p(1-p)$
- Binomial $X \sim B(n, p)$: n independent Bernoulli trials with probability p of each trial
  - e.g., observing the number of derived alleles in a sample with $\frac{n}{2}$ individuals
  - $f(X, \theta) = C_n^X p^X(1-p)^{n-X}$
  - $\mathbb{E}(X) = np, Var(X) = np(1-p)$
  - Most widely used distribution in PopGen
- Normal/Gaussian $X \sim N(\mu, \sigma^2)$: approximation of binomial when $n \to \infty$, $p > 0$

Some important probability distributions II

- (Negative) exponential distribution
- Coalescent: two individuals in a sample with n haplotypes from a population with effective population size N ($2N$ haplotypes)
- The probability of two haplotypes share a common ancestor exactly in generation x is:

  $$f(x, \theta) = (1 - p)^{x-1} p$$

  where $p = \frac{C_n^2}{2N}$
- LD decay: with recombination rate r in each generation, the probability of two loci still keep in LD in generation x is

  $$f(x, \theta) = (1 - r)^{x} L_0$$

  where L_0 is original admixture LD just after admixure
- Both can transform to form of negative exponential distribution

Stochastic process of allele frequency change

- **Markov chain**: present state s_i only depends on the previous state s_{i-1} but not s_{i-2}
- Genetic drift under Wright-Fisher model is a Markov chain
- $$\mathbb{E}[f_{g+1} \mid f_g] = f_g$$
- **Martingale** : expectation of present staet is equal to previous state
- Genetic drift is also a martingale; selection is only a Markov chain but not a martingale

Models of molecular evolutionary processes

- 1.Models based on allele frequency of standing variation (SNP)
  - Using frequency as probability - idea of frequentist:
    $Pr(i) = f_i$, where f_i is allele frequency of locus i
- 1.1.Single locus model: regardless of LD
  - Statistics of the whole genome - average of every loci
  - Drift: $\mathbb{E}(f_{g+1}) = f_g$ (g: generation)
  - Selection: $\mathbb{E}(f_{g+1}) = f_g + s$ (s: selection coefficient)
  - Admixture:
    $$f_i = \sum_k c_k f_i k, \qquad \sum_k c_k = 1$$
    (c_k:ancestry coefficient in ancestral population k)
- 1.2.Two loci model: accounting LD & haplotypes
  - LD of two loci i, j: $D'(i,j) = f_{i,j} - f_i f_j$
  - Haplotype analysis, IBD: more complex model
- 2. Generation of new variation, i.e., mutation
  - Coalescent process, with mean $2N$
  - Accumulated mutations in pairwise individual of a population is $\theta = 2 * 2N\mu = 4N\mu$ (2 lineages)

Two strategies in statistical inference

- 1.Parameter estimation: using data X to infer range (interval estimates) or value (point estimate) of θ
  - Two ways: frequentist (maximum likelihood estimation, MLE) and Bayesian inference
  - Quantified method
  - $f(X \mid \theta)$, finding best θ(s) from multiple/infinite possible θ
- 2.Hypothesis test: constructing a certain distribution $f(X \mid \theta_0)$ conditional on null hypothesis H_0, then testing conditional probability of observing data X and worse results (i.e., p-value)
  - Qualified method
  - Extension of *reductio ad impossible* (proof of contradiction) in probability theory
  - $f(X \mid \theta)$, determing the relationship between H_0 and X by fixing θ_0 under H_0 and calculating conditional probability

Hypothesis test: F_{ST}

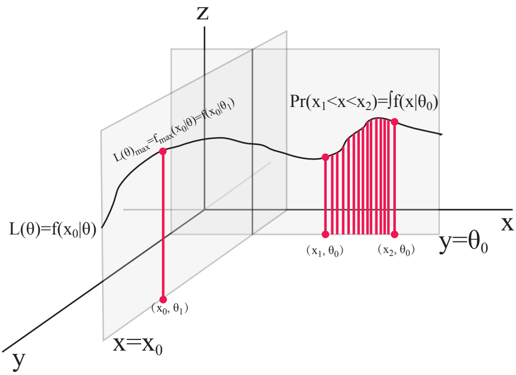
- $$F_{ST} = \frac{\sigma_{intrapop}}{\sigma_{all}}$$
- H_0: no difference between sub-pops, $F_{ST} = 0$

Hypothesis test: D-statistics

- $$D(A, B; C, D) = \mathbb{E}[(p_A - p_B)(p_C - p_D)] = Cov(p_A - p_B, p_C - p_D)$$
- H_0: drift paths $A \to B$ and $C \to D$ are independent $\implies Cov(p_A - p_B, p_C - p_D) = 0$
- Significantly deviated from 0 $\iff$ low probability of observing data under H_0 $\iff$ Shering history & genetic drift between two paths

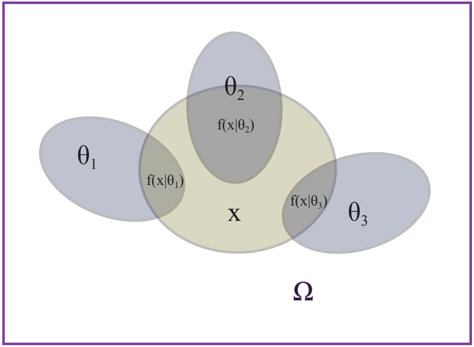
Likelihood and probability

- $L(\theta) = f(X \mid \theta)$
- Probability distribution is function of X, with fixed θ
- Likelihood is function of θ, with fixed X



MLE and likelihood ratio: I

- As frequentists think θ has a single real value, a feasible way to estimate "most probable" θ is finding θ which maximize $L(\theta) = f(X \mid \theta)$. Such a method is **MLE**
- Rationale: finding the point estimate of θ which makes the highest probability of observed data
- Usually using derivation=0 to estimate θ



MLE and likelihood ratio: II

- E.g., in ADMIXTURE, inferring ancestral coefficients c_k by MLE:

$$f_i = \sum_k c_k f_i k, \qquad \sum_k c_k = 1$$

- We can also compare the ratio of likelihood under different parameters $\theta_1, \theta_2, \theta_3 ...$
- E.g., qpGraph, $L(G_1) > L(G_2) > L(G_3)...$

Bayesian inference

- Bayesian inference estimate the interval of θ under posterior distribution (with information given by data)
- Rationale: transforming prior $f(\theta)$, data X, likelihood function into posterior distribution (i.e., distribution of θ under given data X)

$$f(\theta \mid X) = \frac{f(\theta)f(X \mid \theta)}{Pr(X)}$$

- $f(X)$ can be solved by total probability theorem with integration, but this is hard for high dimentional data
- Solution: MCMC/Metropolis-Hastings algorithm, using posterior ratio to avoid calculation of $f(X)$
- A common usage of Bayesian inference is estimation of the interval of coalescent time, as this is "more natural" to be a range than a point value.