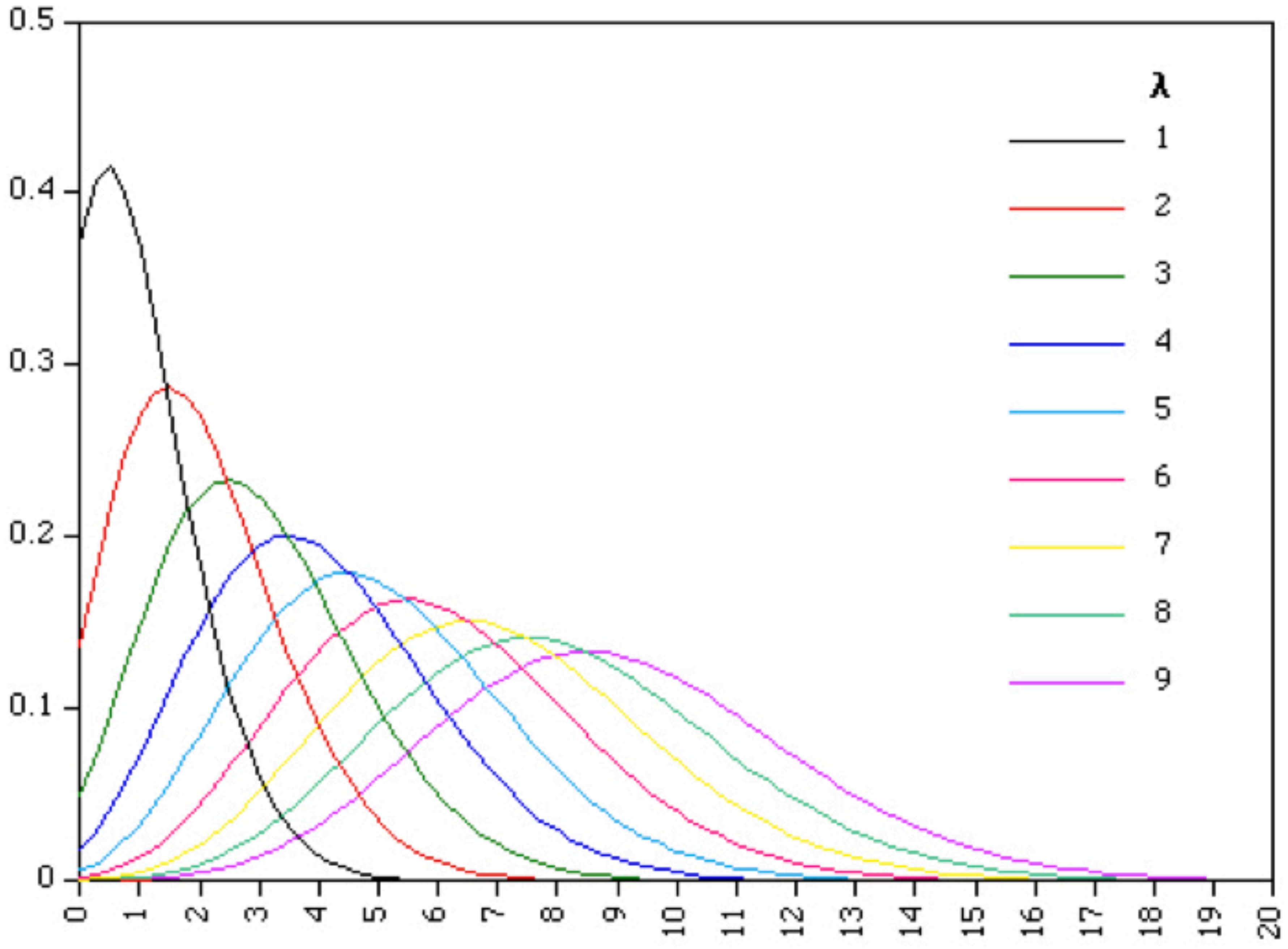
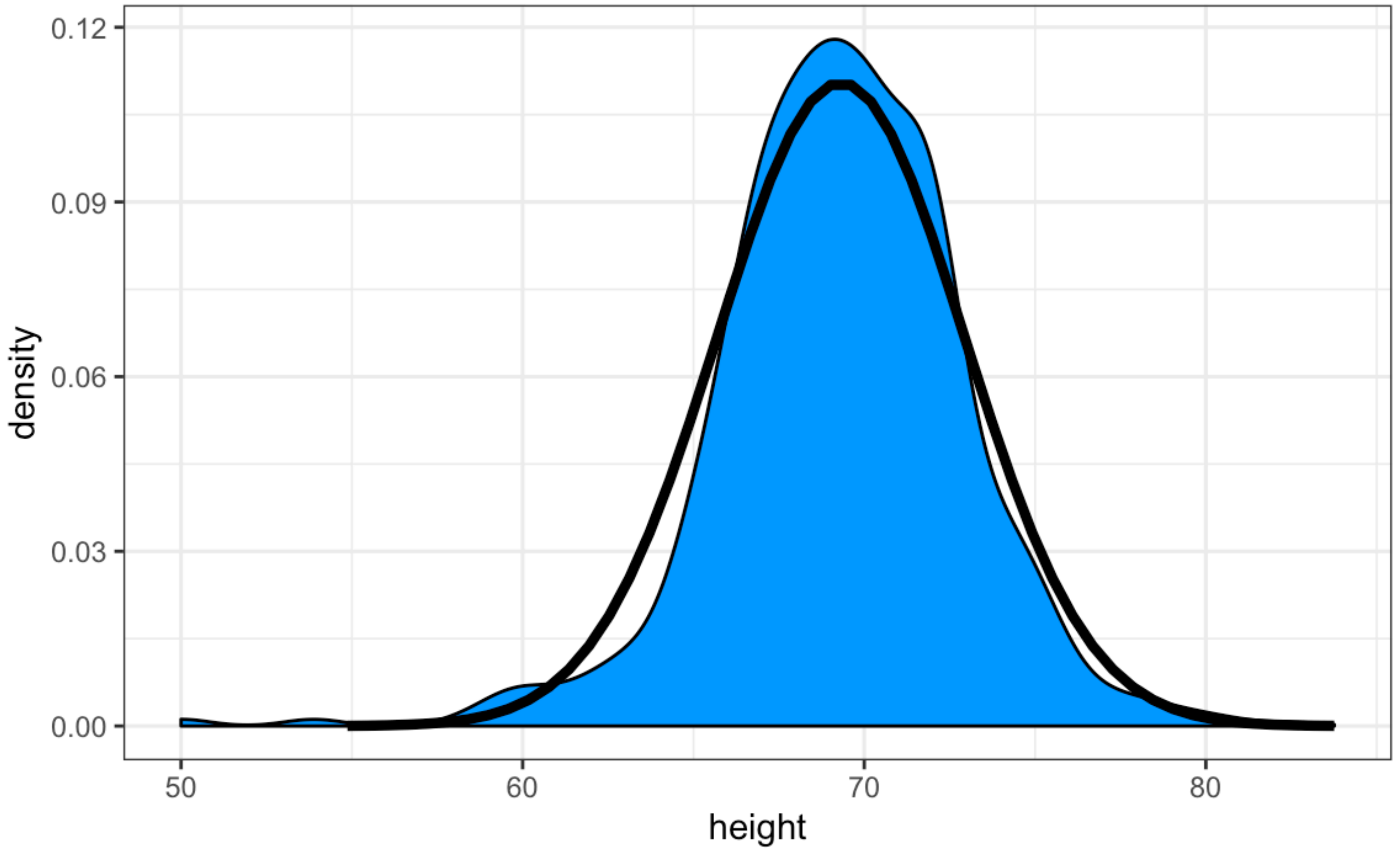
Foundational Statistics

Random Variables and Probability Distributions (Continued)



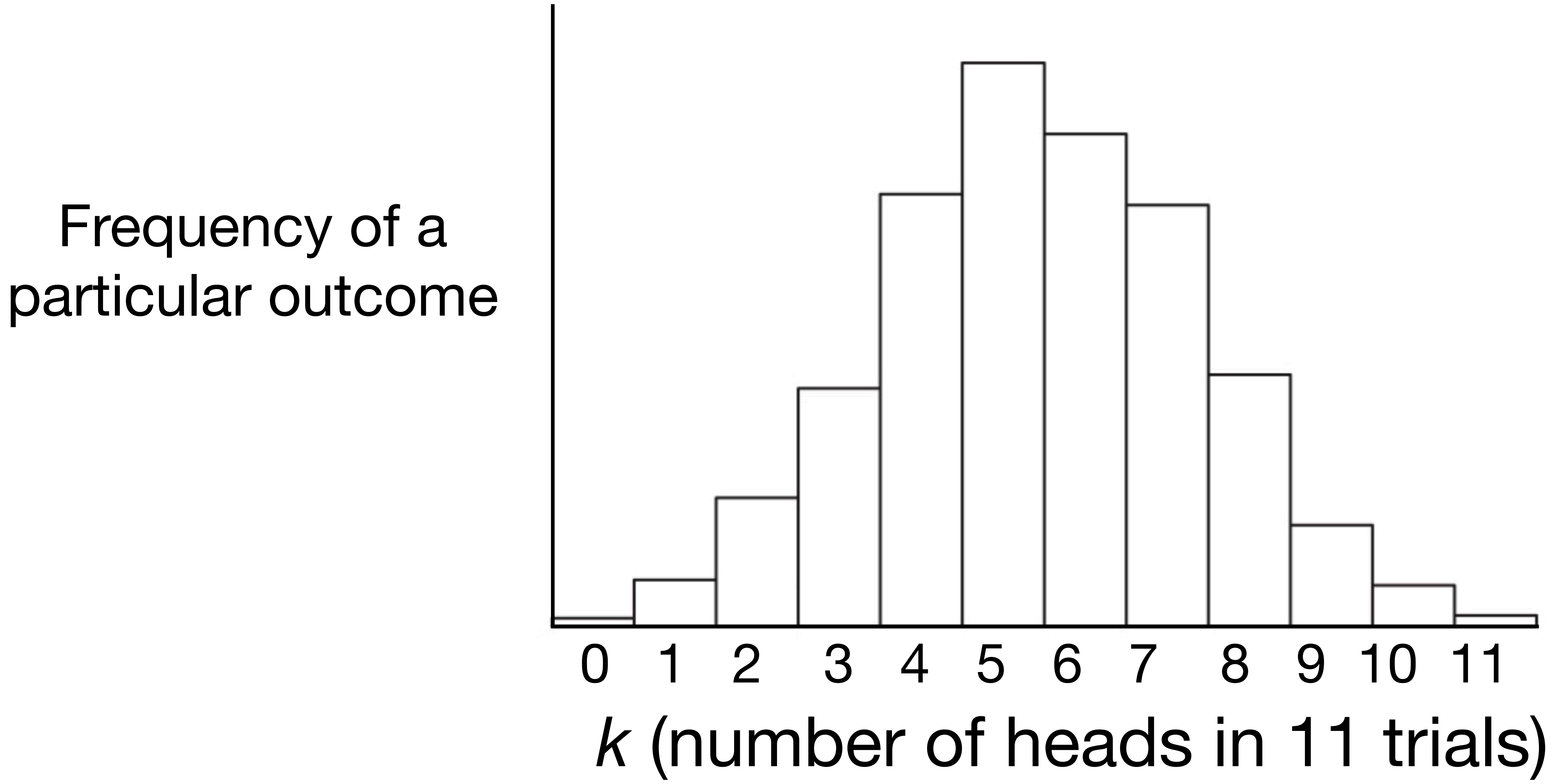


Discrete probability distributions commonly used in the sciences

The Binomial Distribution

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

n = number of trials
k = number of successes in a trial
p = probability of a success



Useful for binary variables
- behavioral choice trials
- presence / absence data
- yes / no survey questions

Discrete probability distributions commonly used in the sciences

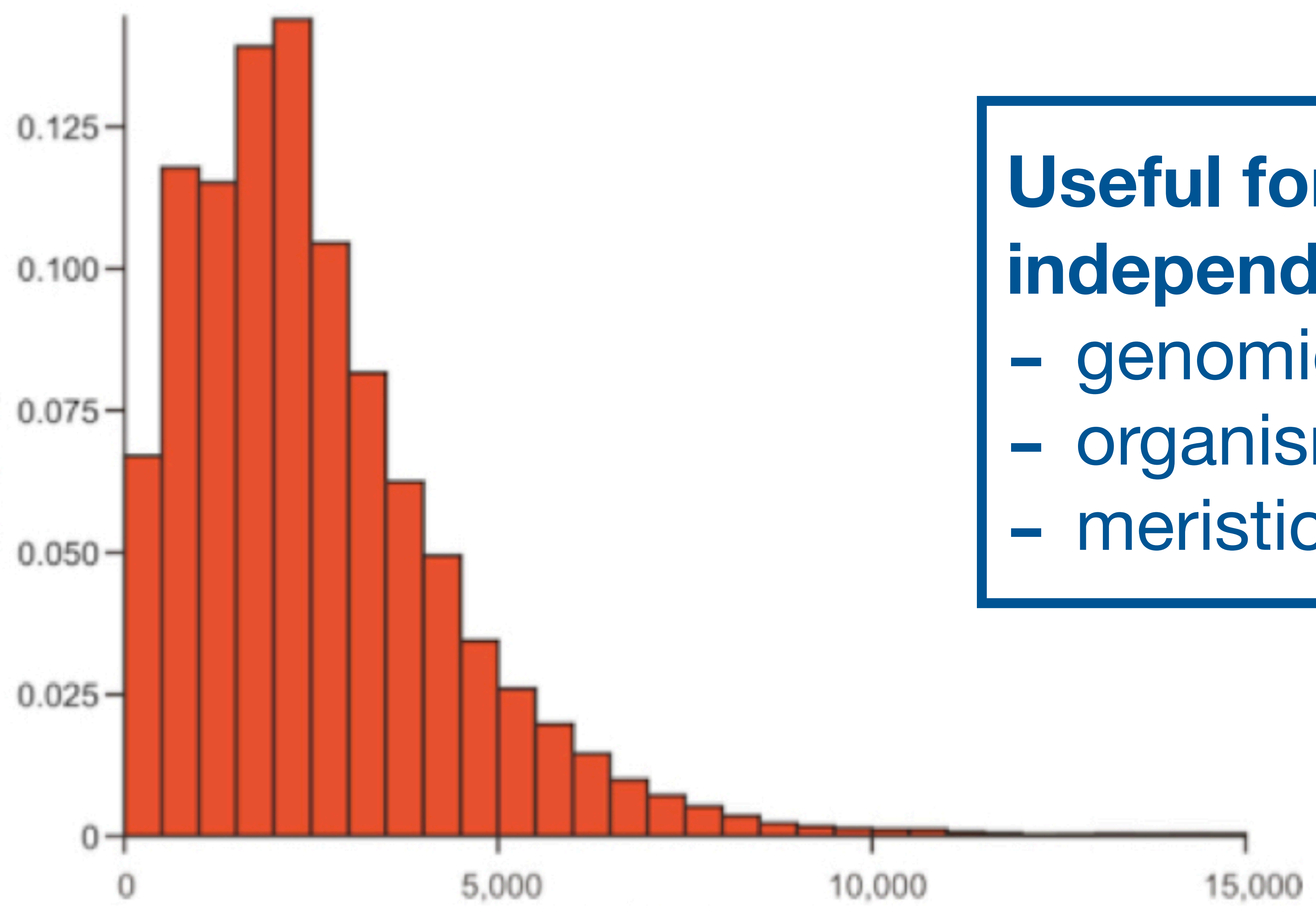
The Poisson Distribution

$$Pr(y = r) = \frac{e^{-\lambda}\lambda^r}{r!}$$

r = count

λ ("lambda") = mean = variance

Frequency of a particular count



Count (bp DNA per gene)

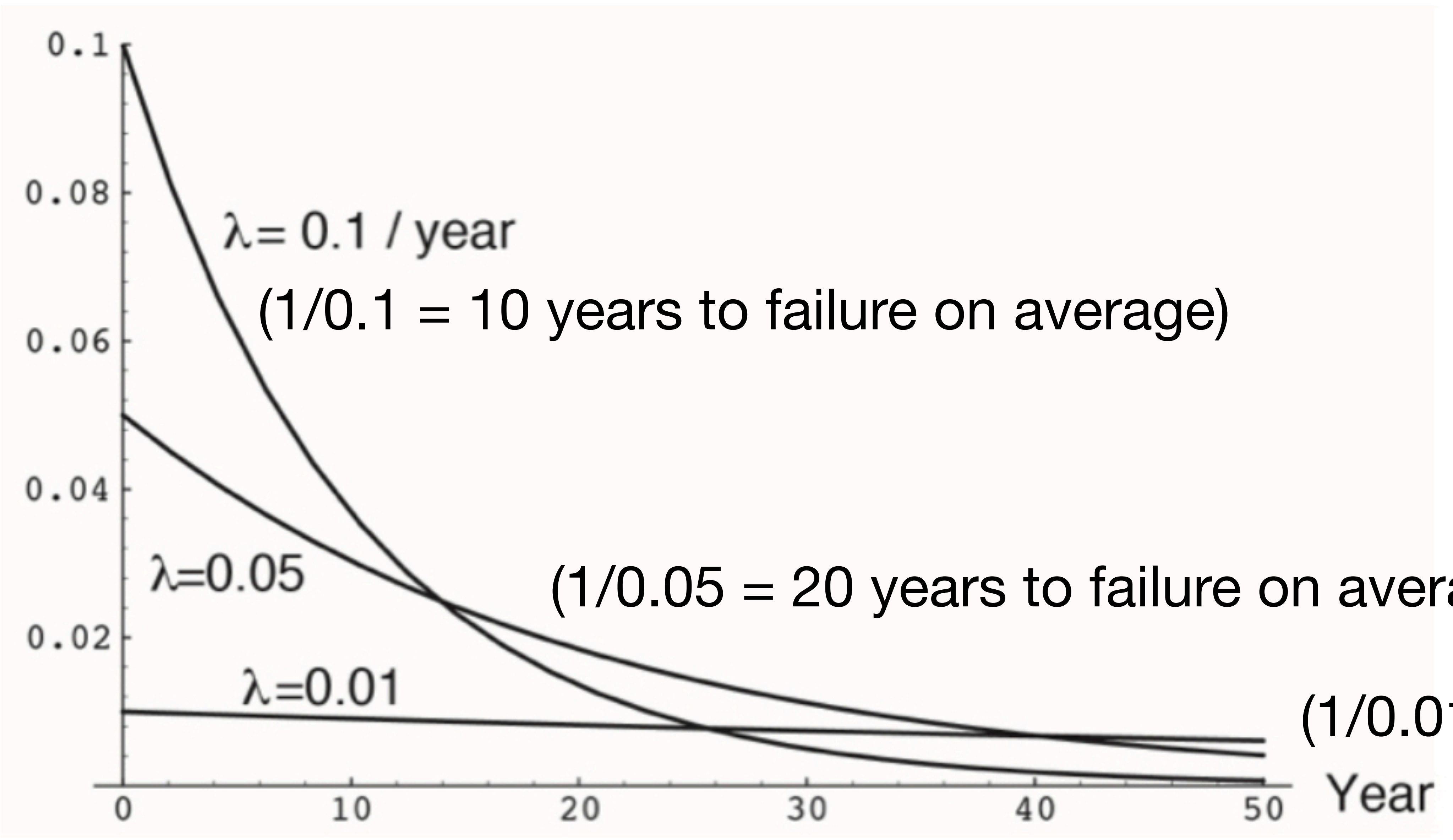
Useful for counts of unlikely, independent events
- genomic features
- organism abundance
- meristic traits

Continuous probability distributions commonly used in the sciences

The Exponential Distribution

$$f(x) = \lambda e^{-\lambda x}$$

λ (“lambda”) = rate



Useful for time-to-occurrence studies

- survival analysis
- radioactive decay
- materials failure analysis

Continuous probability distributions commonly used in the sciences

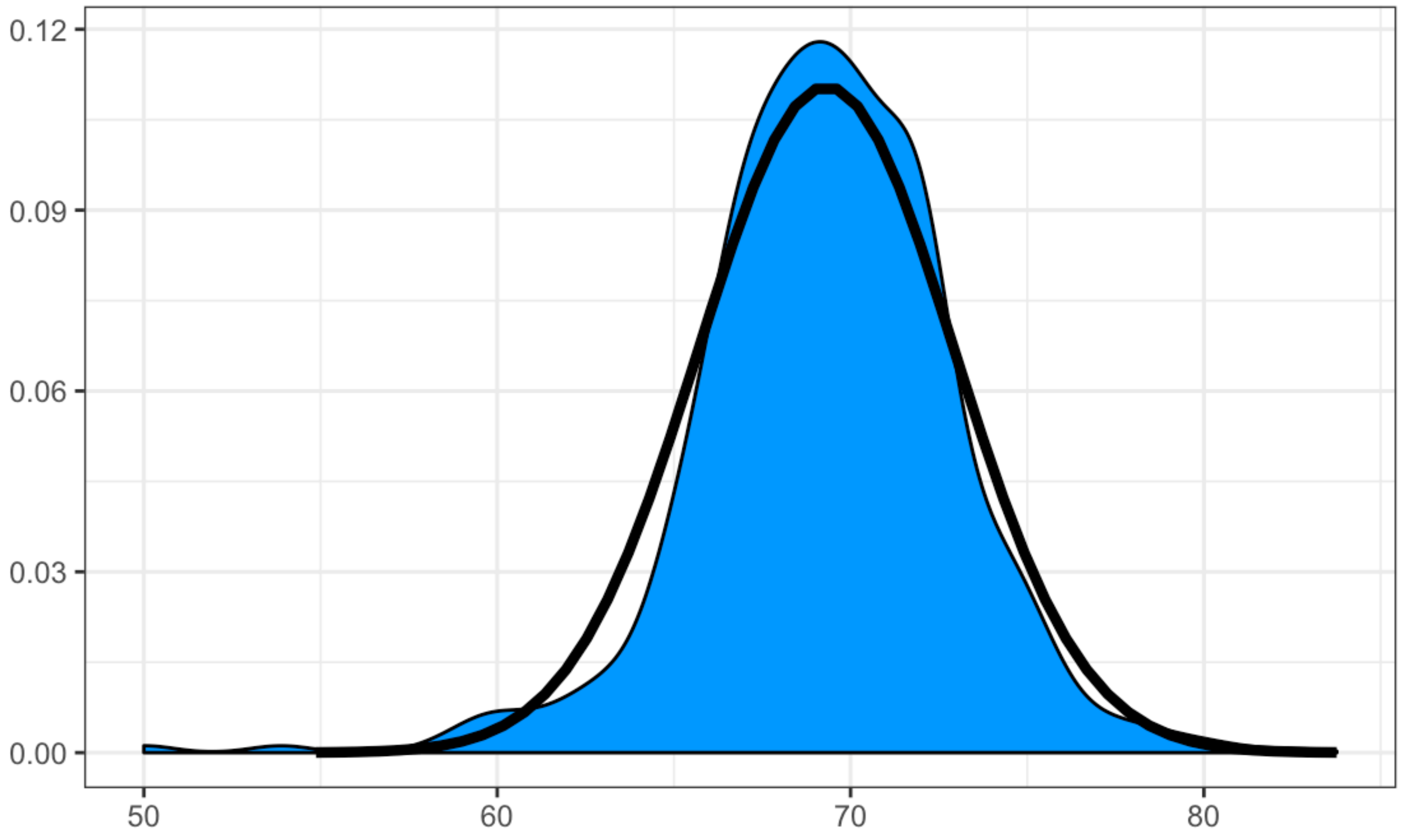
The Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ (“mu”) = mean

σ (“sigma”) = standard deviation

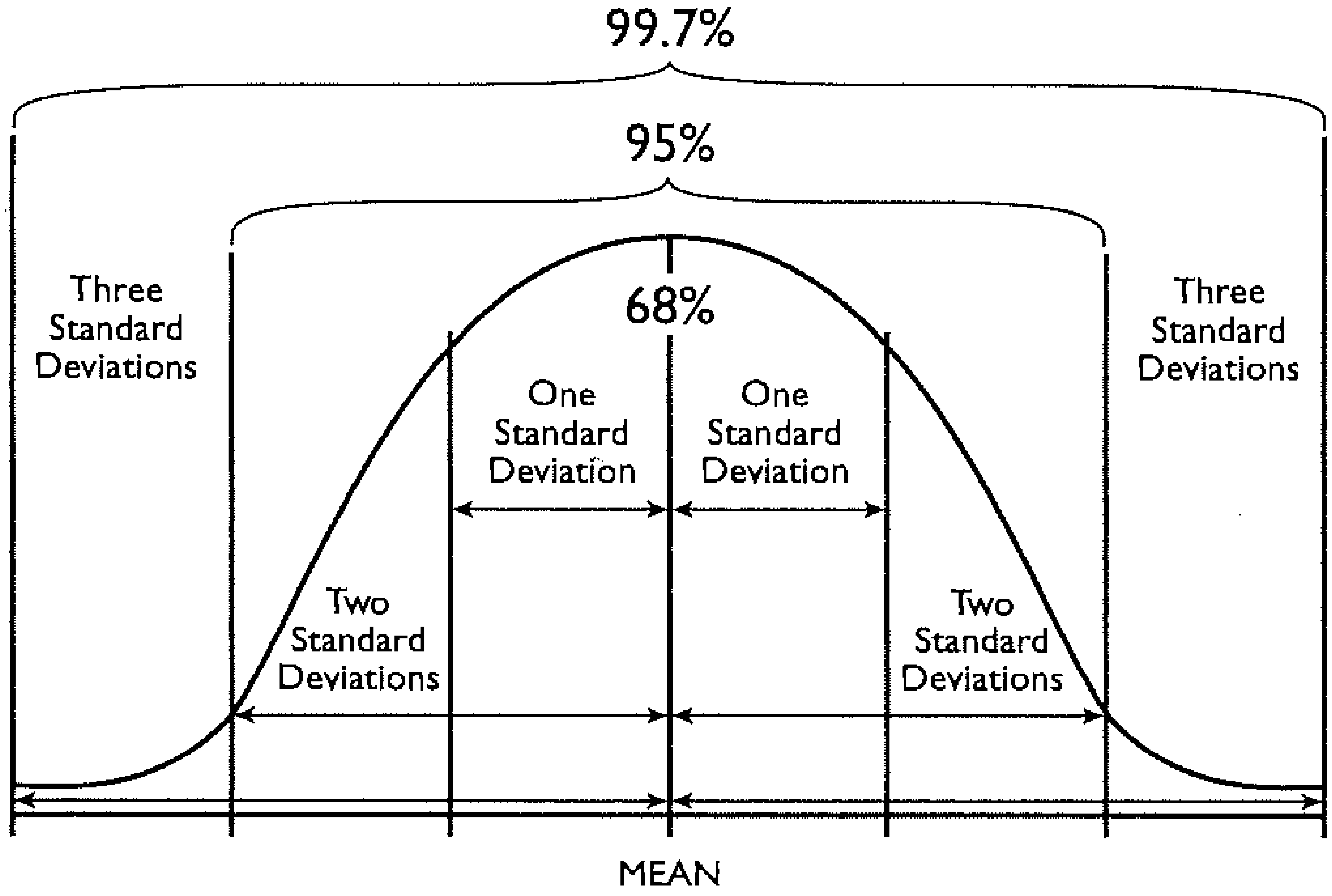
Frequency of a particular value



(Height)

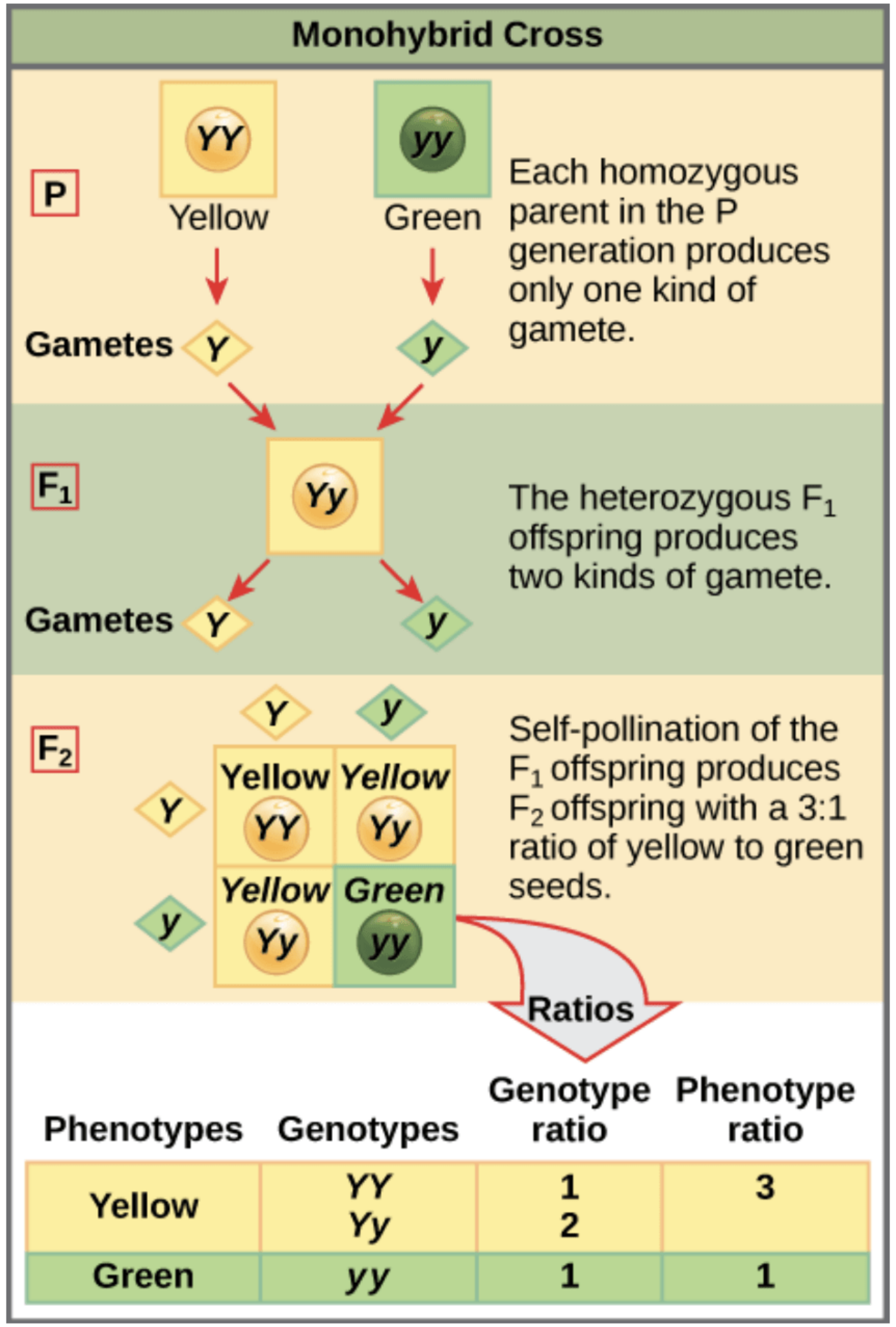
Useful for many random, continuous variables
- complex traits in biology
- physical processes
- social science metrics
- measurement error

The Normal Distribution



Mean = Median = Mode

A historical controversy

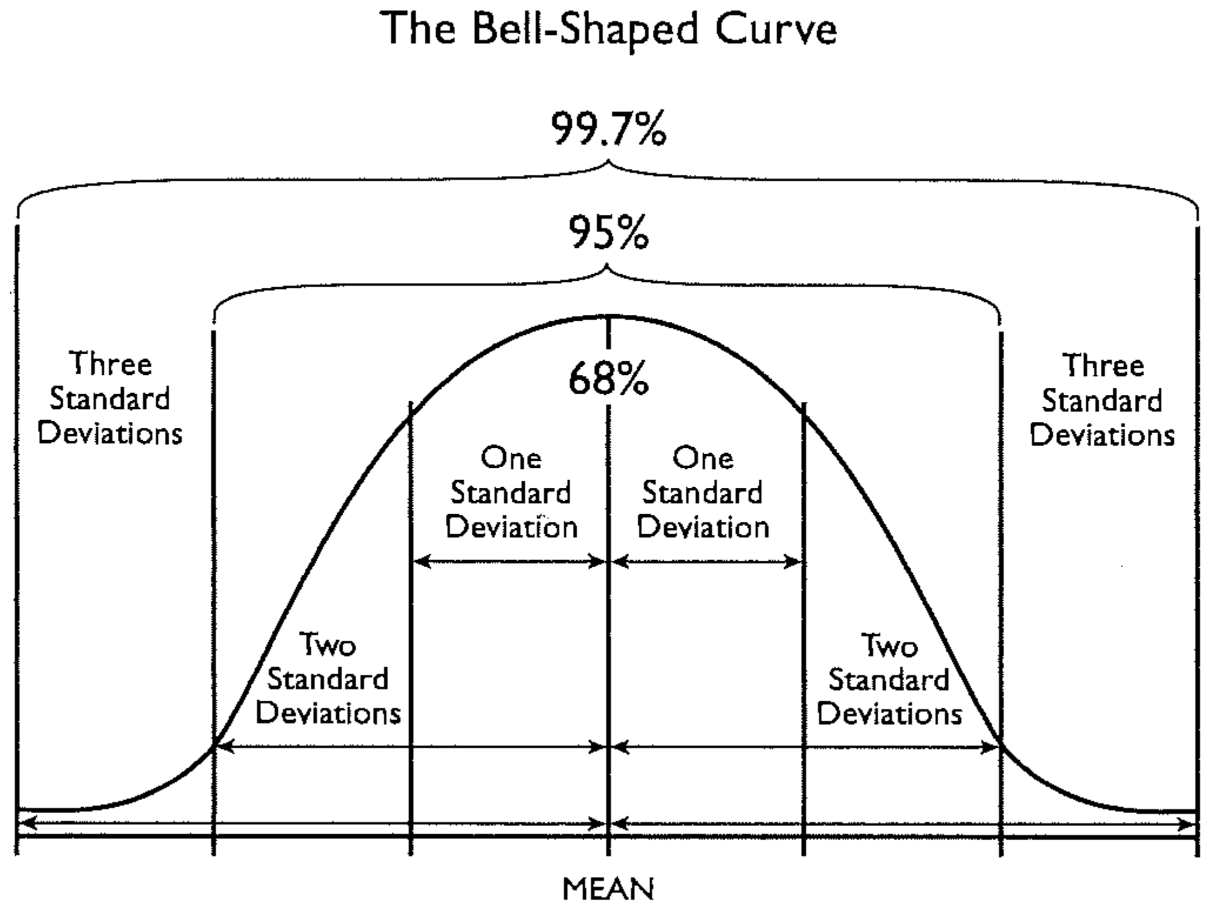


OpenStax Biology

The "Mendelians"

discrete genetic factors -> <u>discrete phenotypes</u>

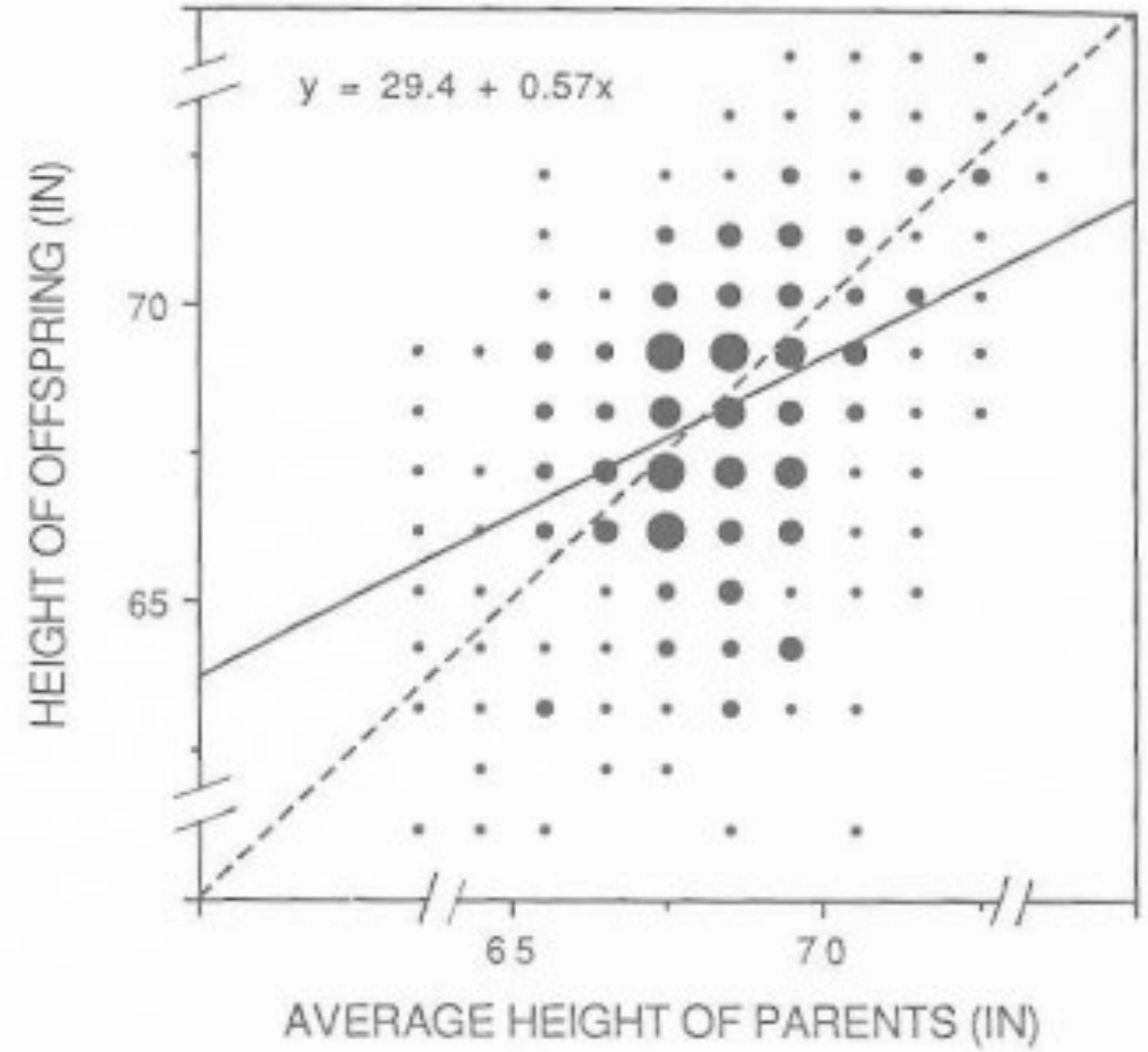
evolution: <u>large steps</u>



? genetic factors -> <u>continuous phenotypes</u>

evolution: <u>gradual</u>

The "Biometricians"



A historical controversy (reconciled)

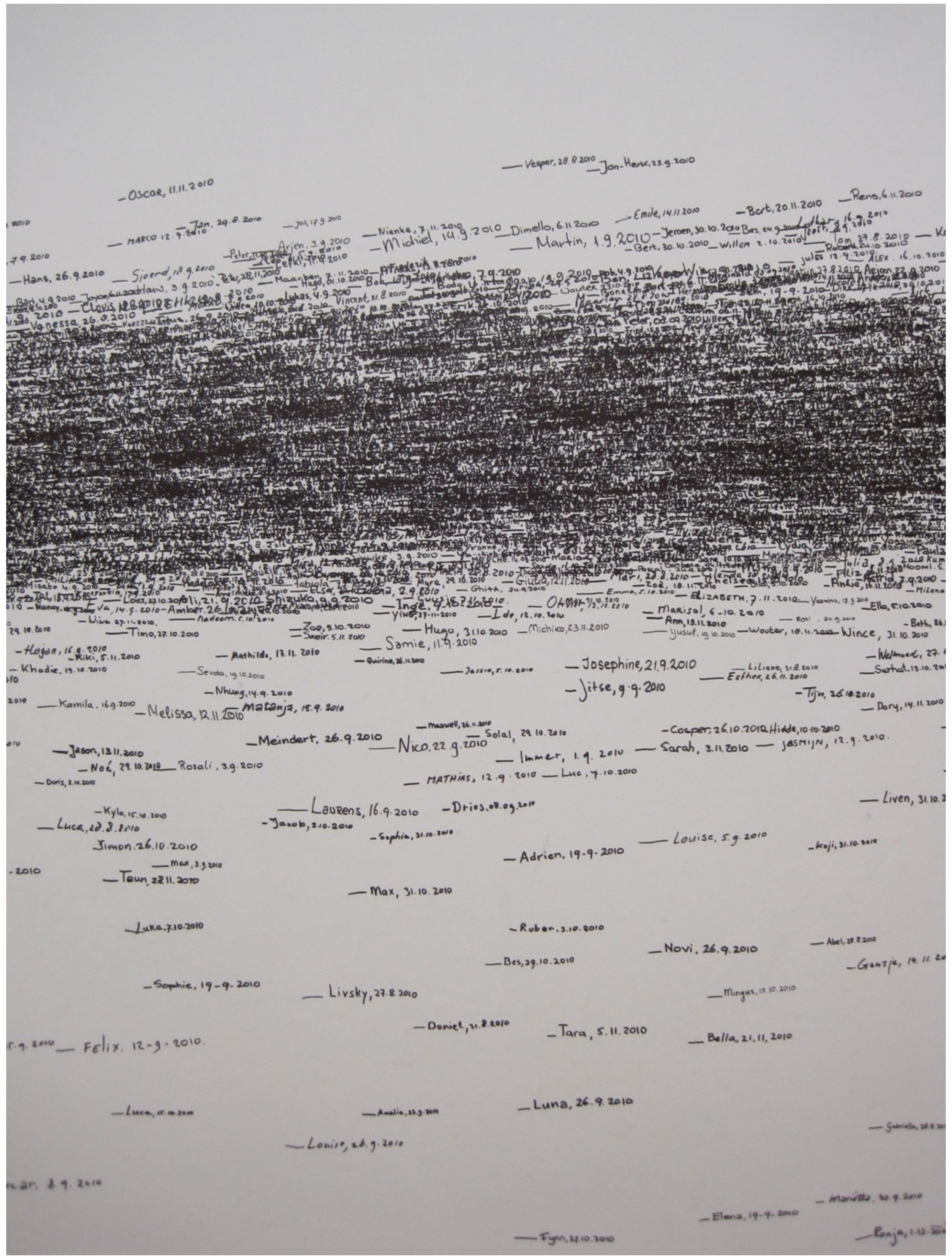
Moving toward synthesis: "The multiple-factor hypothesis"

1 (diallelic) locus -> **3 genotypes**: AA Aa aa

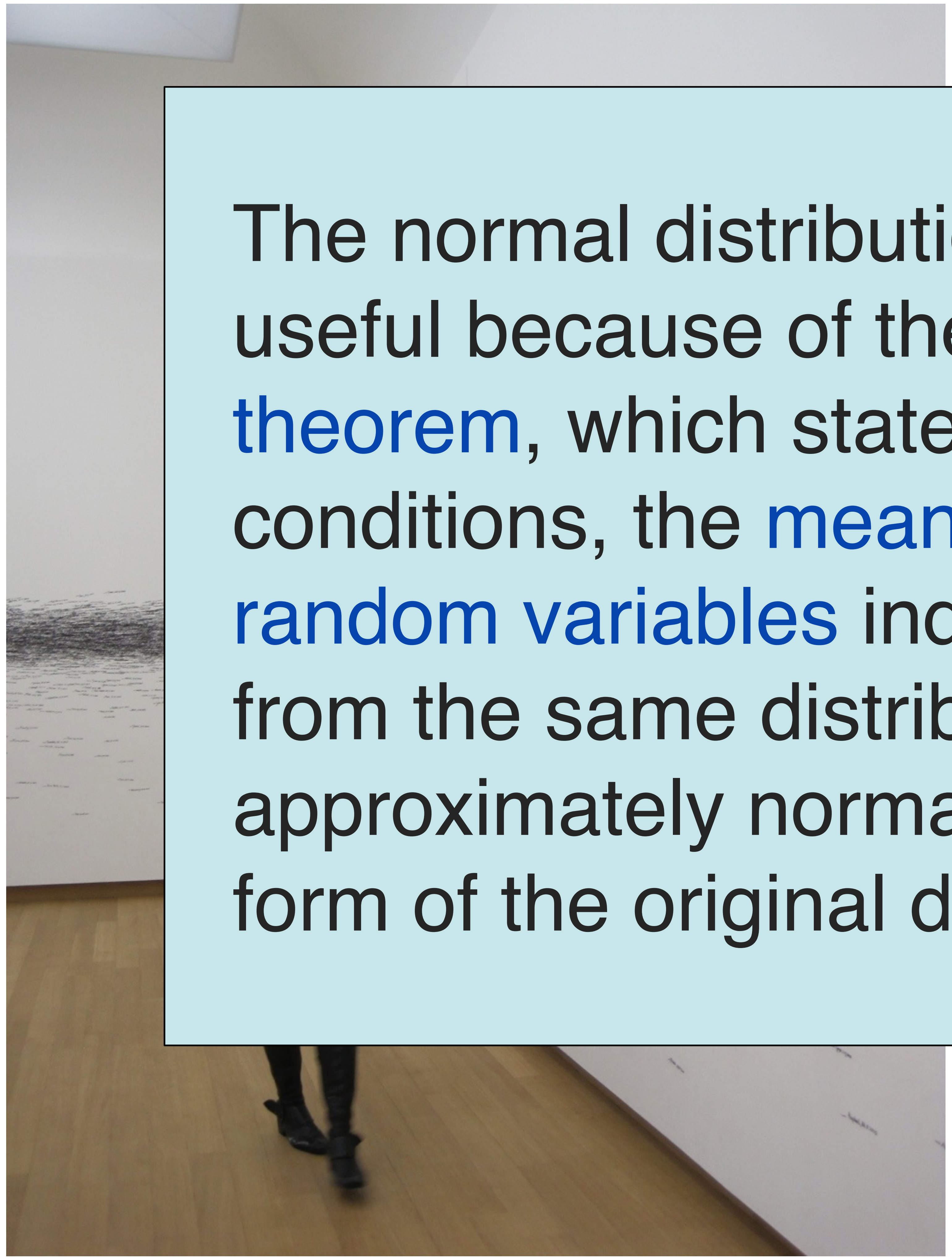
10 (diallelic) loci -> $3^{10} \approx$ **60,000 genotypes**

Also: Non-genetic (environmental) factors add continuous variation

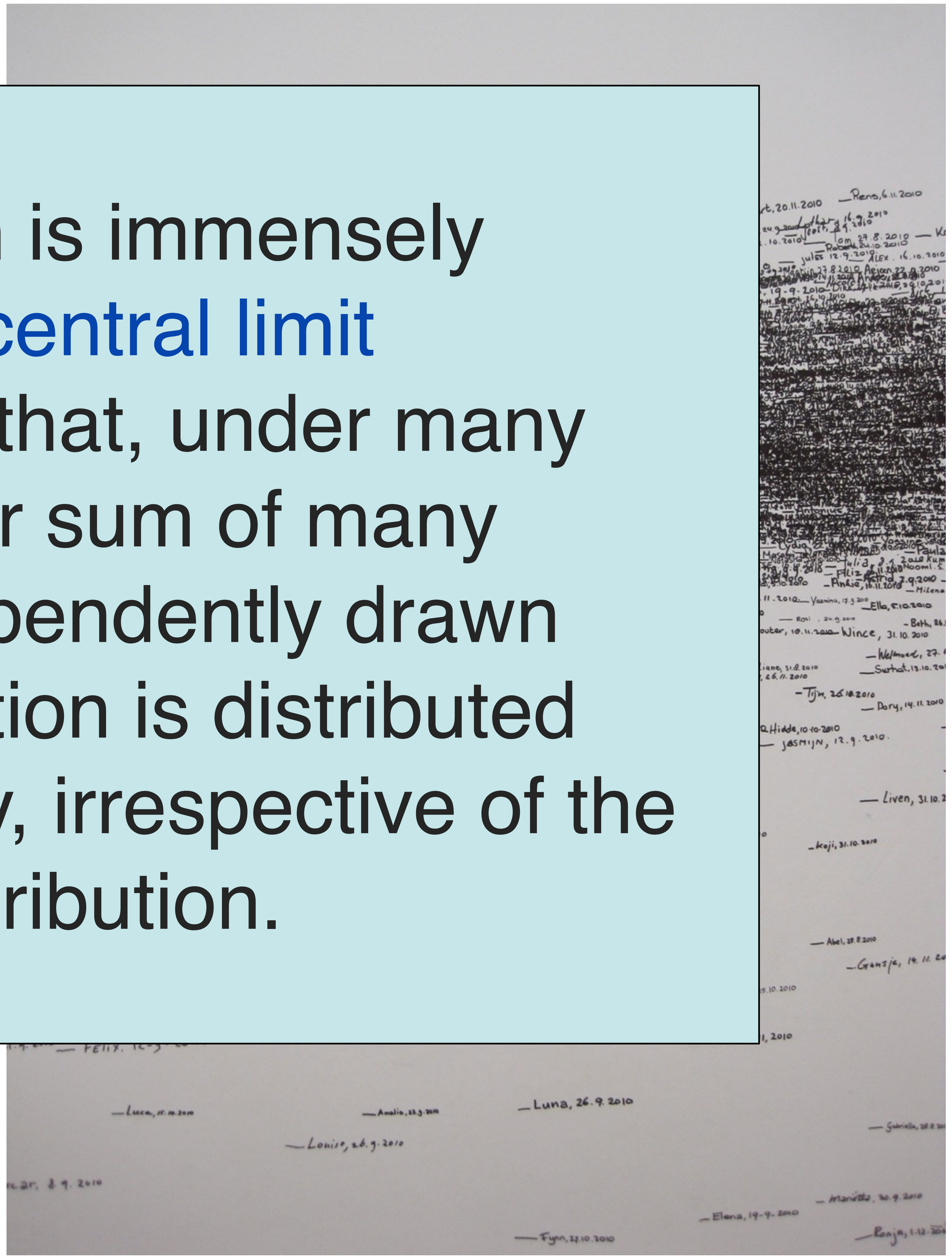
A depiction of human height variation



The CLT is useful when thinking about random samples from a variety of prob. distributions



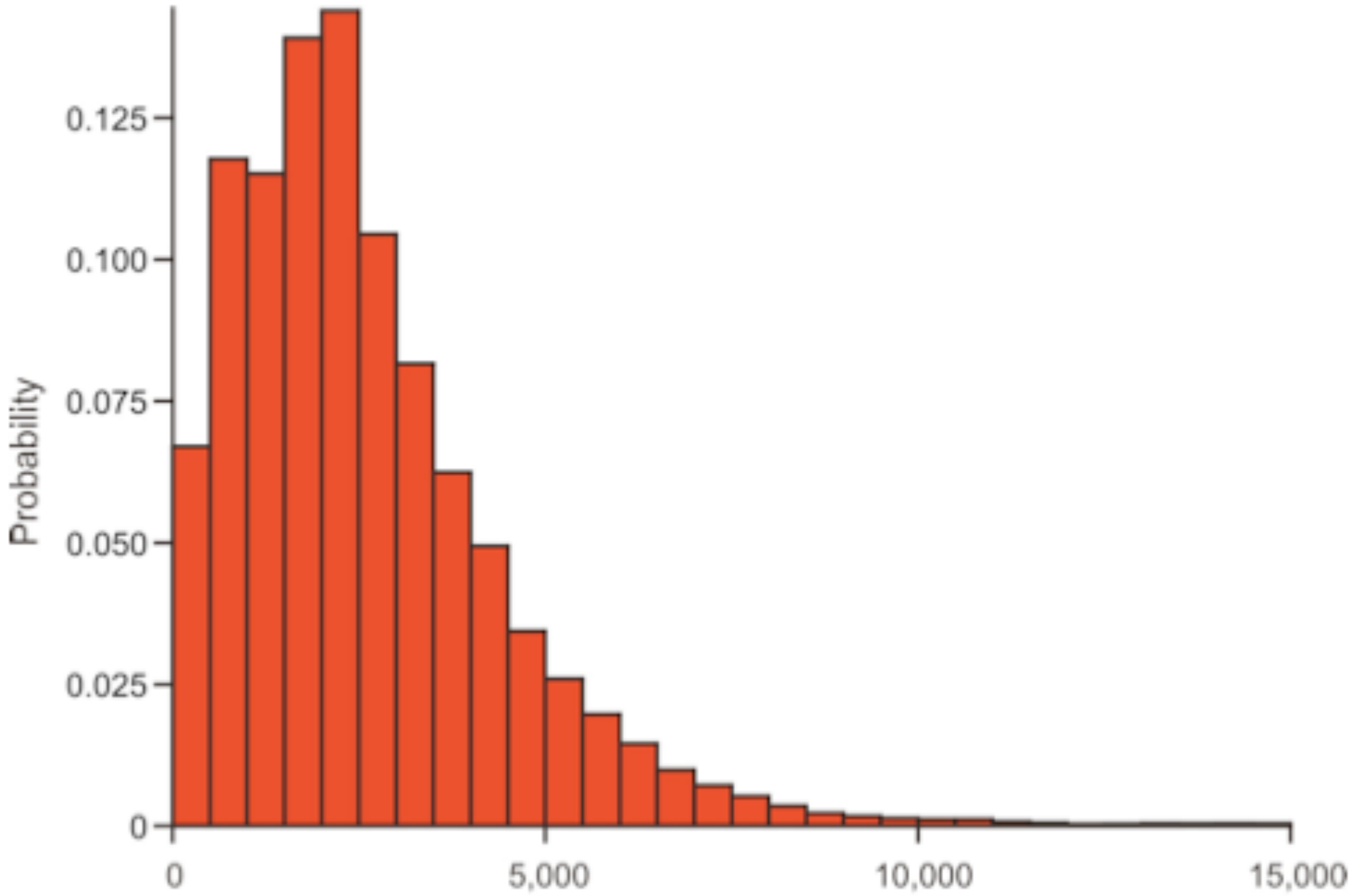
The normal distribution is immensely useful because of the central limit theorem, which states that, under many conditions, the mean or sum of many random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution.



The CLT and sampling distributions

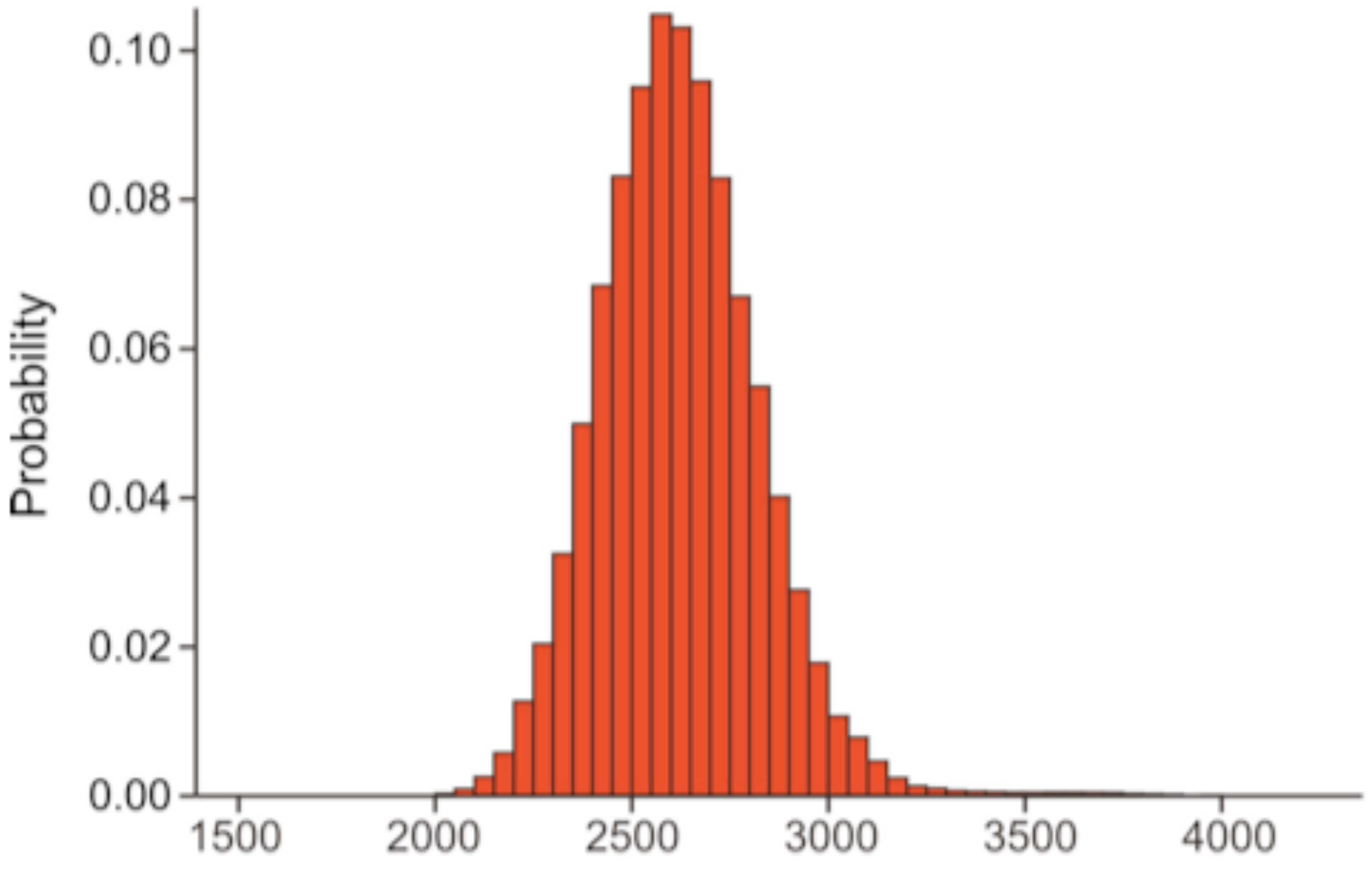
Sampling variation around a parameter = *often* normal

Random Variable Distribution (Poisson)



bp DNA per gene

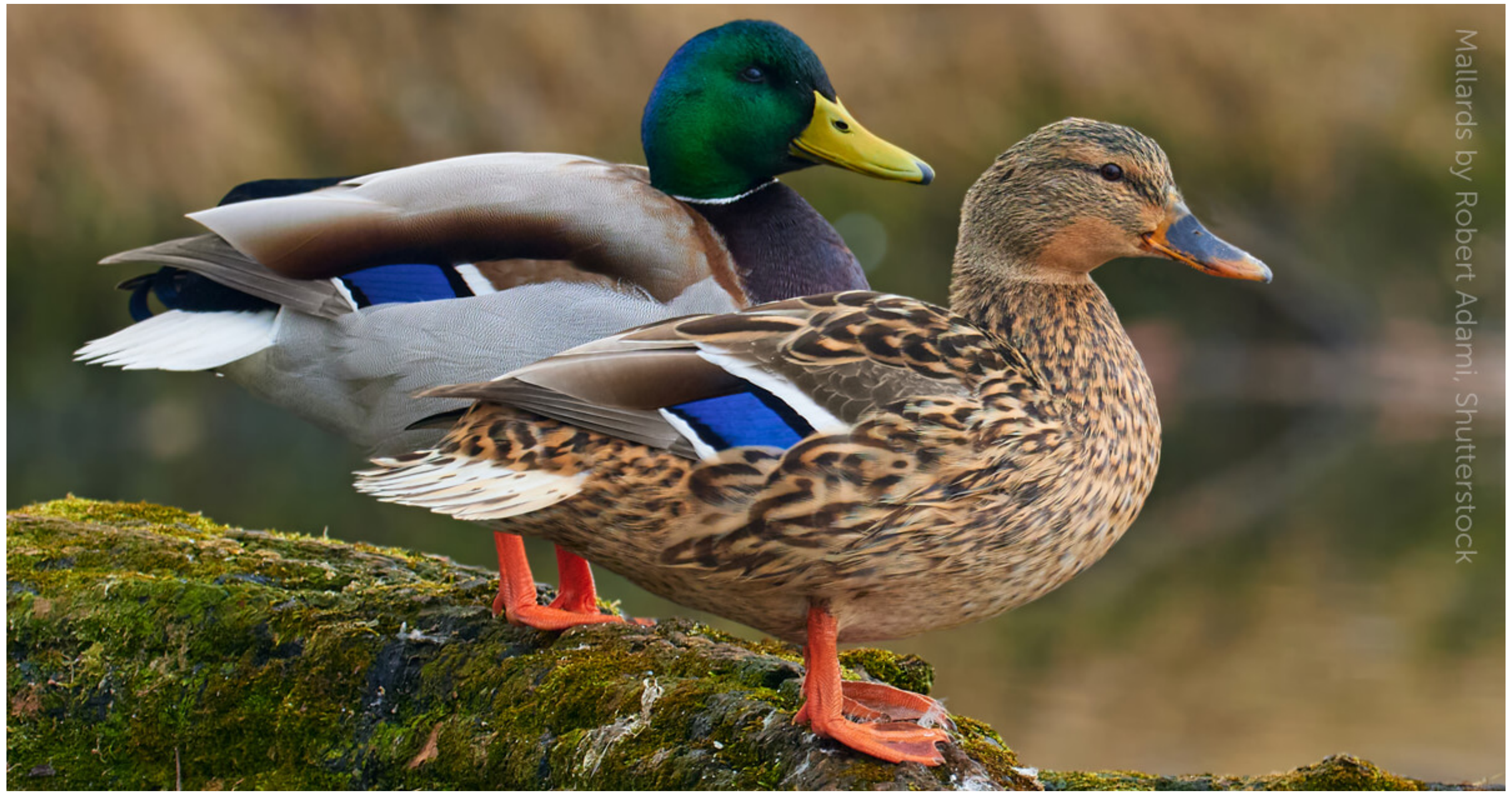
Sampling Distribution (Normal)



Mean bp from a sample

Simulating the CLT for a quantitative trait in R

Can we simulate a ≈ Normal distribution for 500 mallard duck bill length measurements, from discrete genotypes at 5 loci?



What "ingredients" do we need for our simulation?