Module 3A : Thinking in Distributions

Building block for Hypothesis Testing

Agenda:

- Major distributions:
  - **Discrete Distributions**
    - Bernoulli
    - Binomial
    - Poisson
    - Hypergeometric
    - Uniform (disc. or cont.)
  - **Continuous Distributions**
    - Normal
    - Exponential
    - Gamma
- Interactive simulations
- **Central Limit Theorem**
  - **Sampling Distribution of the mean**

- Probabilities build up to distributions
  - Gives information about location and spread of the data
- There are standard distributions that explain common biological phenomenon

- **Discrete Distributions**
  - Bernoulli
  - Binomial
  - Poisson
  - Hypergeometric

- **Continuous Distributions**
  - Normal
  - Uniform
  - Exponential
  - Gamma

Websites for simulations:

1. https://seeing-theory.brown.edu/probability-distributions/index.htm
2. https://probstats.org/

Bernoulli Distribution

Bernoulli Trials:

- ***Random*** *process with* ***only two mutually exclusive outcomes***
  - Coin toss: heads versus tails; Contest: Win or lose
  - General: one is called a success, one is called a failure

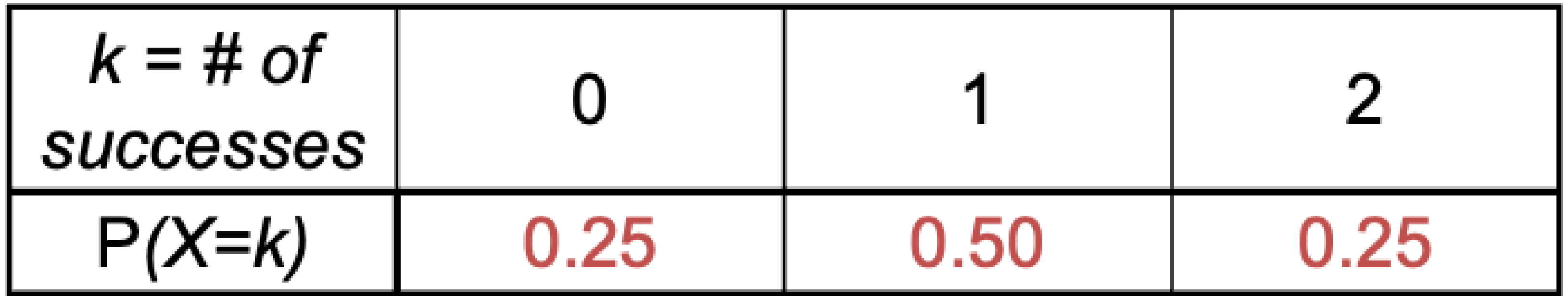
- *The probability,* ***p****, of success is the same in every trial*

- *The trials are* ***independent-*** *the outcome of any particular trial has no influence on the results of any other trial*

- *Visualization:* ***https://probstats.org/bernoulli.html***

Binomial Distribution

- **Binomial Random Variable**
    - Repeat a Bernoulli trial, with probability of success ***p***, to get a Binomial Random Variable
    - **X** is the number of successes in a fixed number, ***n***, of repeated Bernoulli trials
        - Example: P(*X* = *k*), where X represents the number of heads in two coin flips so *k* = 0,1,2



k = # of *successes*	0	1	2
P(*X*=*k*)	0.25	0.50	0.25

TT(0.5*0.5), HT(0.5*0.5), TH(0.5*0.5), HH(0.5*0.5)

Visualization: https://probstats.org/binomial.html

- **Binomial Distribution:**
  - Describes the probability of a given number of ‘**successes**’, which have a ***p*** probability, from a fixed number of independent trials, ***n***

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

- **The Binomial coefficient:**

  It counts all the unique unordered sequences of getting *k* successes in *n* trials.
  i.e. how many ways are there of getting k successes?

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

Where n! = n x (n-1) x (n-2) x ... x1

Also: **0! = 1 and 1! = 1**

- The Binomial coefficient:

$$\boxed{\binom{n}{k} = \frac{n!}{k!(n-k)!}}$$

Example: How many ways can 2 letters be chosen from the set {A B C D}?

Example: What is the probability of getting exactly the following pattern (2 successes and 3 failures):
F F S F S

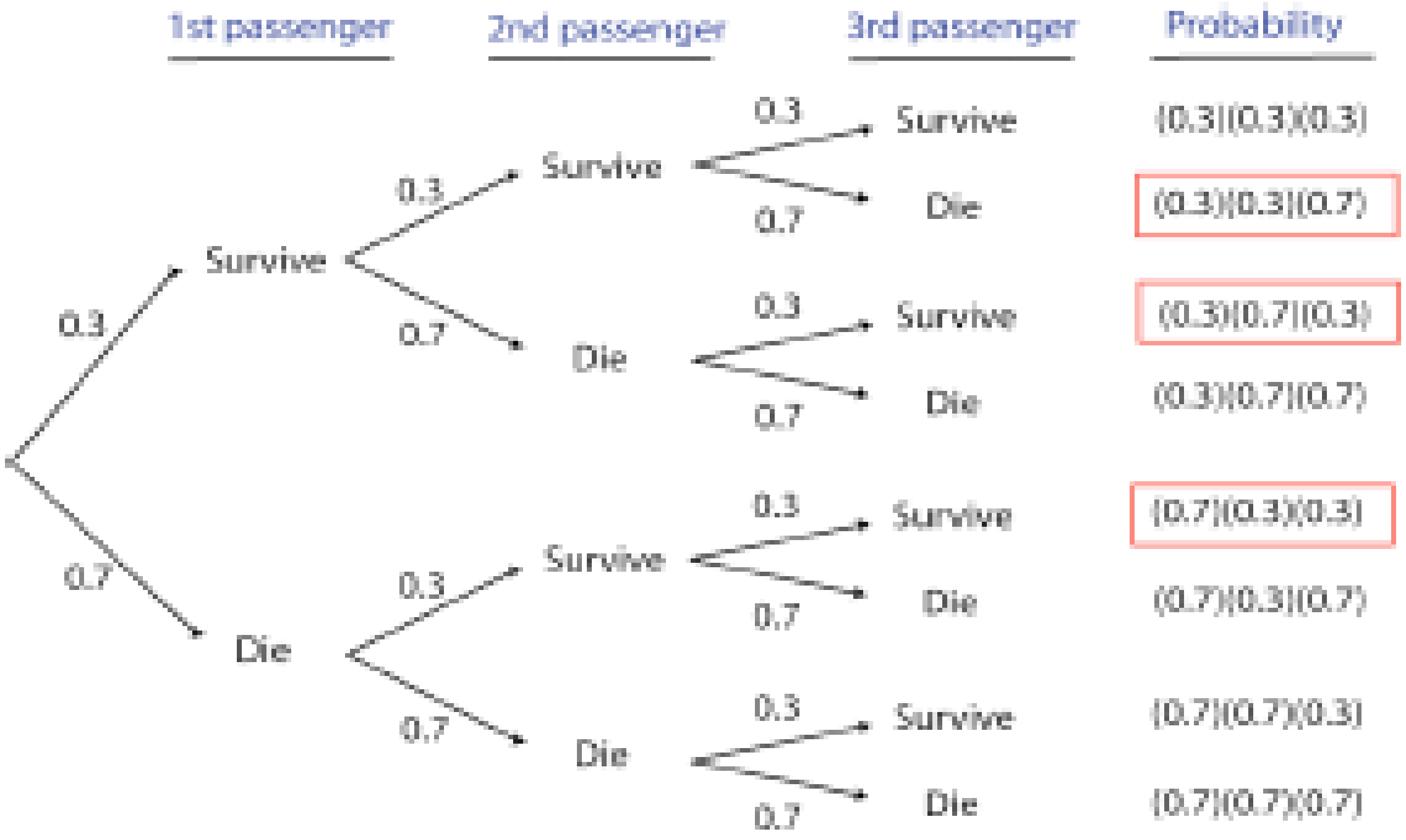
Binomial Distribution

Example: **2092** passengers on the titanic; **654** survived

P(surviving) = 654/2092 = **0.3**

Question : *What is the probability that 2 out of 3 randomly chosen passengers survived?*

Answer: The hard way…



Properties of the Binomial Distribution:

A binomial random variable has values that are the number of successes

The long way of demonstrating the mean and standard deviation:

If 40% of brand A widgets have a particular defect, and I buy 5 of these widgets, what is the expected number of defective widgets that I now own?

Use https://probstats.org/binomial.html with n=5, p=0.40 to visualize the distribution…..

Properties of the Binomial Distribution:

A binomial random variable has values that are the number of successes

ANSWER (hard way):

Outcome:	0 widgets	1 widget	2 widget	3 widget	4 widget	5 widget
Probability:	0.07776	0.25920	0.34560	0.23040	0.07680	0.01024
Random Variable:	0	1	2	3	4	5

$$\boxed{\bar{X} = 0(0.07776) + 1(0.25920) + 2(0.34560) + 3(0.23040) + 4(0.07680) + 5(0.01024) = 2}$$

This is the same answer as would be obtained by simply multiplying the probability of “success” times the number of cases….

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$

Poisson Distribution

The *Poisson distribution* is a discrete distribution modeling the number of times an event occurs in a time interval, given that the average number of events occurring in the interval is known. You can think of the Poisson distribution as a special case of the Binomial distribution but with a **really large number of intervals and a really small probability of success in any given one interval** (in more math-y speak: n approaches infinity and p approaches 0). You only need to know the mean number of an event – it is useful to not need to know the exact number of intervals or where the events happened to describe, for instance, how mutations are distributed along genealogies. There is an accessible explanation of how to derive the Poisson distribution from the binomial distribution here (warning: there are limits involved): https://medium.com/@andrew.chamberlain/deriving-the-poisson-distribution-from-the-binomial-distribution-840cc1668239

$$P(x;\mu) = \frac{e^{-\mu}\mu^{x}}{x!}$$

Mean = Variance = μ

What is the probability of having 110 novel mutations (mutations that neither of your parents have) if the mean mutation rate of the human genome is 115?

https://probstats.org/poisson.html

Hypergeometric Distribution

- This is used in tag-and-release programs.
- Basis for one-tailed version of Fisher’s exact test (we will see this later)

https://probstats.org/hypgeom.html

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

https://en.wikipedia.org/wiki/Hypergeometric_distribution

Uniform Distribution

https://probstats.org/uniform_discrete.html

- used as an “information-less” prior in Bayes’ Formula

Formula: f(x) = $\frac{\mathbf{1}}{\boldsymbol{max-min}}$

Mean: $\frac{\boldsymbol{max+min}}{\mathbf{2}}$

Variance: $\frac{(\boldsymbol{max+min})^{\mathbf{2}}}{\mathbf{12}}$

Question: A chromosome is 2 Morgans in length (recall that one Morgan corresponds to the genomic distance of a mean of one crossover event). Is a mean observed position of the crossover is at 1 Morgan with a variance of 1/2 consistent with a uniform distribution of crossover events along the 2 Morgan chromosome?

Normal Distribution

- Bedrock of inferential statistics
- Approximate the Binomial Distribution (n >30).
- Phenomenon that result from many additive small effects processes are normally distributed --- this is very common in biological processes (Single mutation Mendelian traits are the exception rather than the rule)
- **Central limit theorem:** means of samples of random variables from other distributions (not normal distributions) can approach a normal distribution as then sample size increases.

Formula: $\boldsymbol{f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}}$

Mean: $\boldsymbol{\mu}$

Variance: $\boldsymbol{\sigma^2}$

(The complicated denominator ensures that the distribution integrates to 1.)

- https://probstats.org/normal.html

Exponential Distribution

https://probstats.org/exponential.html

Formula: $\boldsymbol{f(x) = \alpha e^{-\alpha x}}$
Mean: $\boldsymbol{\frac{1}{\alpha}}$
Variance: $\boldsymbol{\frac{1}{\alpha^2}}$

Question: A bee is foraging and stops at flowers at a constant rate,$\alpha = 0.05$ per meter. What is the mean distance travelled between flowers?

Central Limit Theorem

- CLT allows us to assume that any sampling distribution of the mean is normally distributed…. Even if the random variables are from a highly skewed distribution (you will need to increase n if you are sampling from a highly non-normal distribution)

- Note that the Binomial Distribution involves summing the outcome of n independent Bernoulli trials so, as predicted by the CTL, it is roughly normally distributed.

- You can build an intuition for this by drawing the difference between the allele distribution of Aa*Aa (4 squares) mating and AaBb*AaBb mating (16 squares) and AaBbCc*AaBbCc (64 squares).

The sum of n independent and identically distributed random variables tends toward a normal distribution as n → ∞

The Correlation between Relatives on the Supposition of Mendelian Inheritance (1918). R.A. Fisher

Fisher's 1918 paper introduces the term *variance,* and it demonstrates that the *discrete* inherited traits proposed by Mendel could give rise to traits that displayed *continuous* variation, i.e. human height. This profound insight allowed Mendelian genetics to explain Darwinian natural selection and laid the groundwork for the last century of modern biology.

Fisher's paper is challenging to read but, for our interest, we can think about it as a quincunx simulator, a simple demonstration of how to get a normal distribution from multiple discrete alleles.

Francis Galton (1822-1911) was the first to note that many biological traits (height, weight etc.) followed a normal distribution. He invented the quincunx (also called the "Galton Machine" or "Bean Machine") that gave insight into **"regression to mediocrity".**

https://www.mathsisfun.com/data/quincunx.html

1. We take any continuous trait (height is a popular trait for this type of example since the alleles of >10,000 gene variants contribute to it and only a small fraction of the variation in height is currently explained by all those genes).

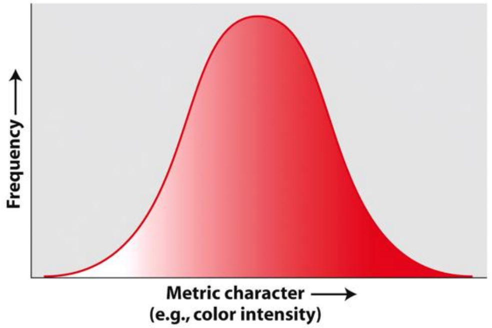


Figure 3-14
Introduction to Genetic Analysis, Tenth Edition
© 2012 W. H. Freeman and Company

2. We can see that a simple heterozygous cross (let's visualize it as AaBb X AaBb) can give us 5 different classes of the trait, if having a dominant allele endows the trait with a dose of 1 (whatever that dose translates to in terms of traits): 0 doses (aabb), 1 dose (Aabb, or aaBb), 2 doses (AaBb or AAbb or aaBB), 3 doses (AABb or AaBB) or 4 doses (AABB).

Female Gametes/Male Gametes	AB	Ab	aB	ab
AB	AB/AB	AB/Ab	AB/aB	AB/ab
Ab	Ab/AB	Ab/Ab	Ab/aB	Ab/ab
aB	aB/AB	aB/Ab	aB/aB	aB/ab
ab	ab/AB	ab/Ab	ab/aB	ab/ab

Female Gametes/Male Gametes	AB	Ab	aB	ab
AB	4 doses	3 doses	3 doses	2 doses
Ab	3 doses	2 doses	2 doses	1 dose
aB	3 doses	2 doses	2 doses	1 dose
ab	2 doses	1 dose	1 dose	0 doses

1. We take any continuous trait (height is a popular trait for this type of example since the alleles of >700 genes contribute to it and only a small fraction of the variation in height is currently explained by all those genes).
2. We can see that a simple heterozygous cross (let's visualize it as AaBb X AaBb) can give us 5 different classes of the trait, if having a dominant allele endows the trait with a dose of 1 (whatever that dose translates to in terms of traits): 0 doses (aabb), 1 dose (Aabb, or aaBb), 2 doses (AaBb or AAbb or aaBB), 3 doses (AABb or AaBB) or 4 doses (AABB).
3. Finally, the tables above can result in the normal approximation to the binomial distribution:

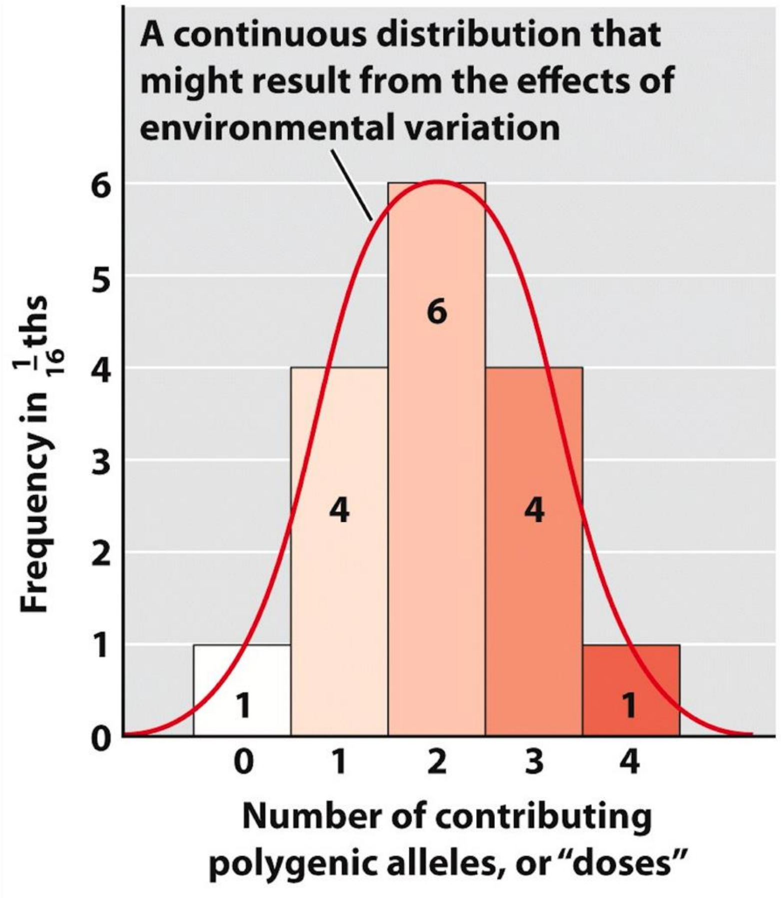
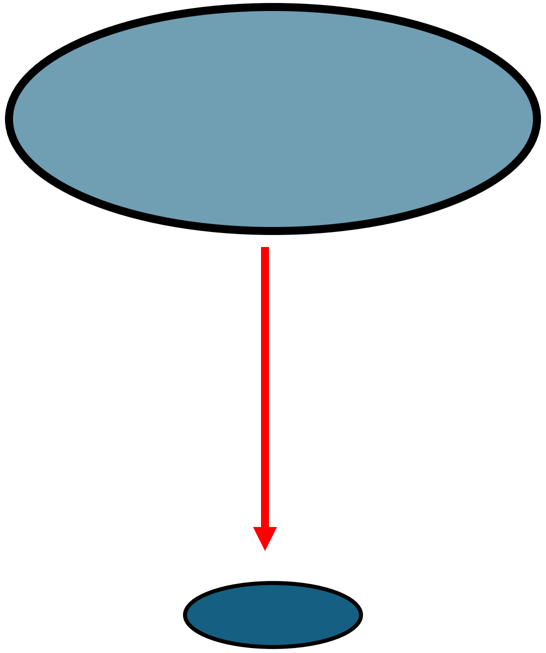


Figure 3-16
Introduction to Genetic Analysis, Tenth Edition
© 2012 W. H. Freeman and Company

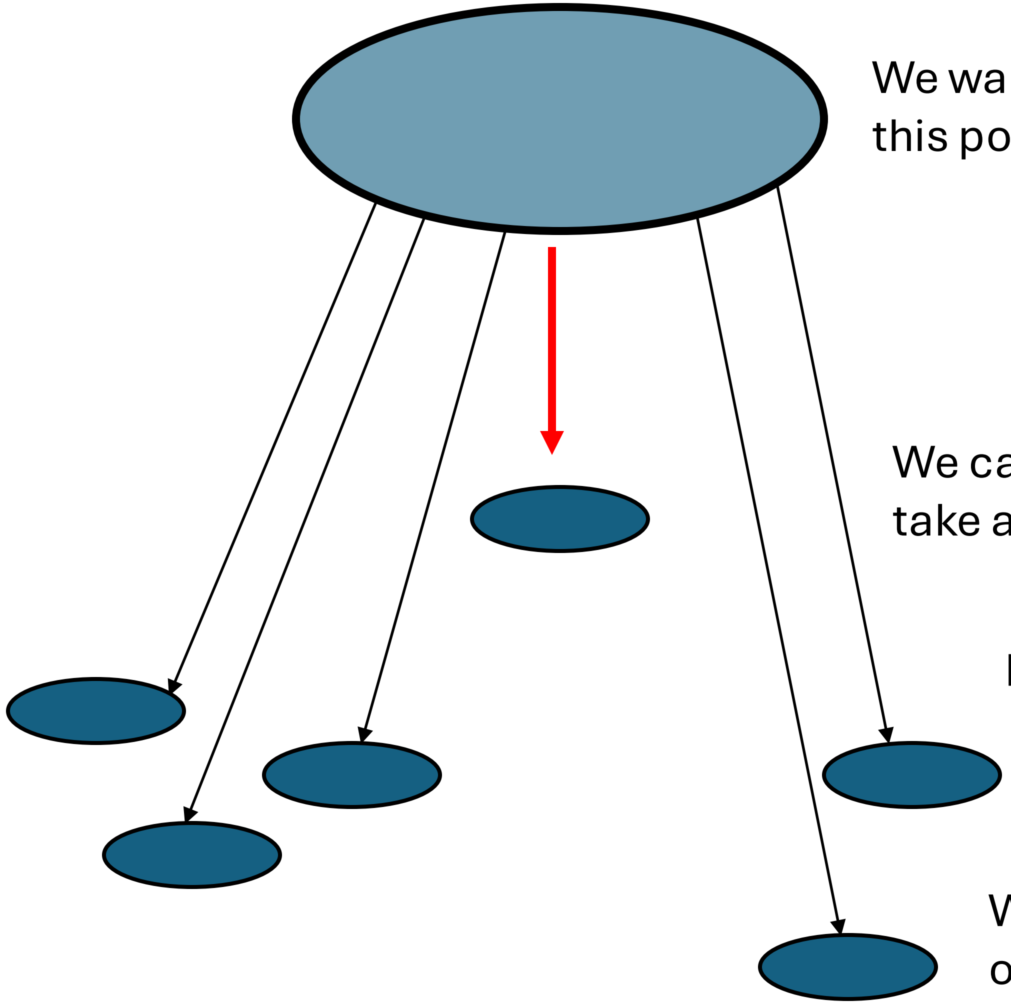
Sampling matters:



We want to know something about this population

We can’t measure everyone, so we take a sample

But! the sample doesn’t necessarily have the same properties as the population due to chance errors.

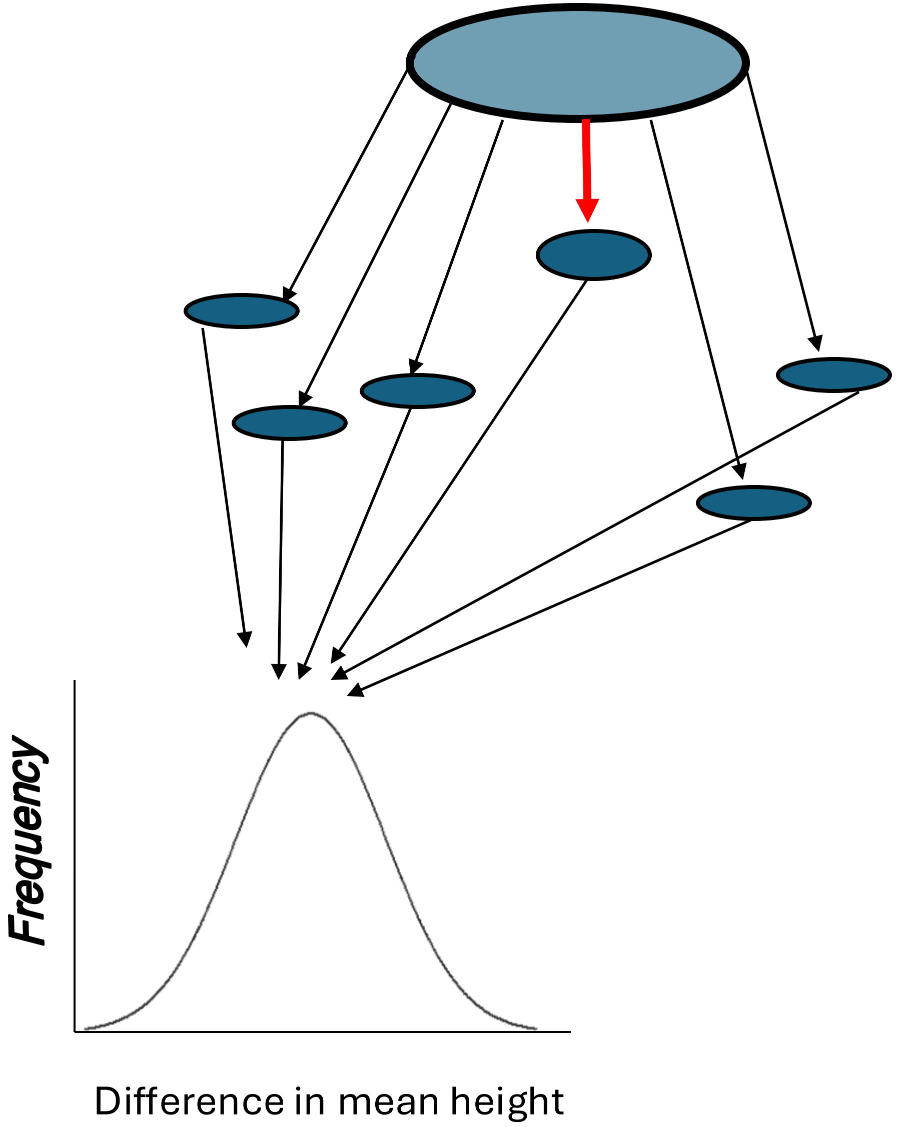


We want to know something about this population

We can’t measure everyone, so we take a sample

How ‘*good*’ is the sample?

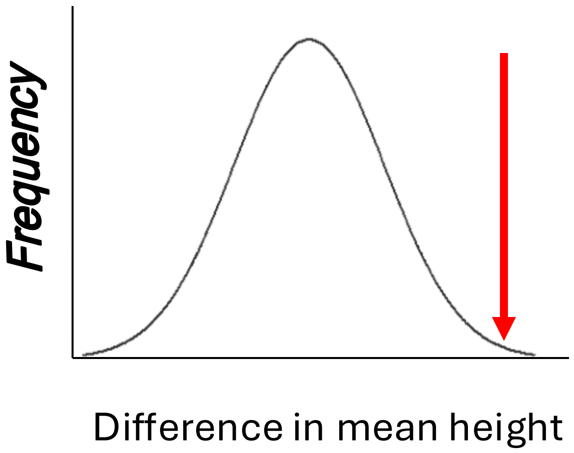
We imagine taking an infinite number of samples from the ***null distribution***



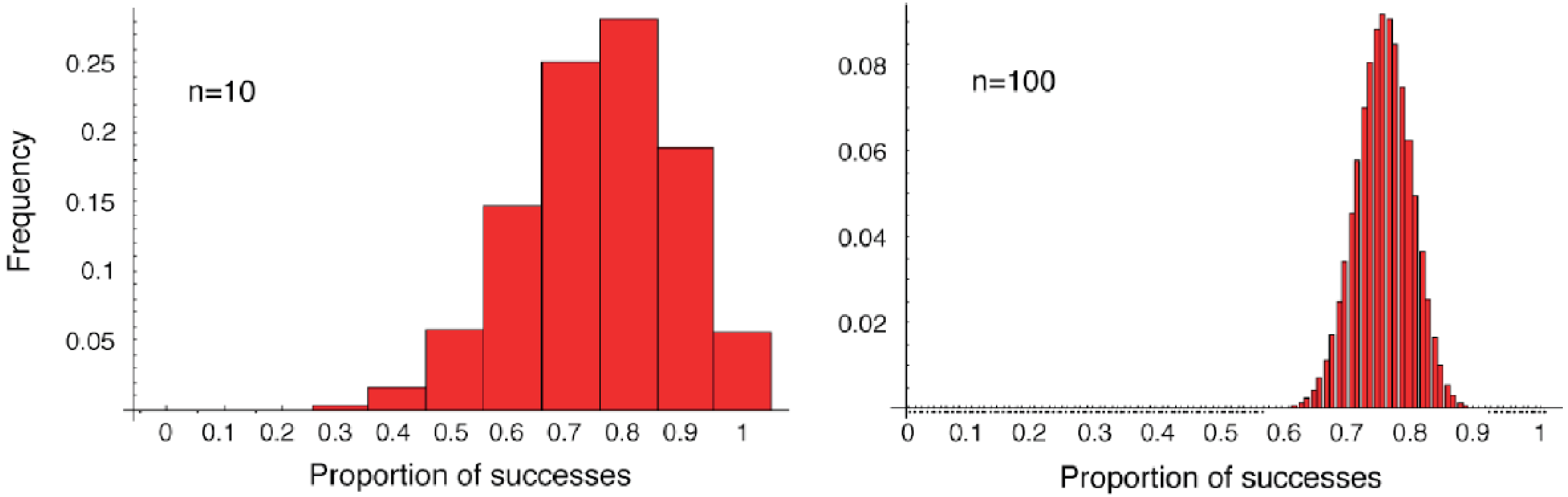
- We want to know something about this population
- We can’t measure everyone, so we take a sample

How ‘good ’is the sample?

We imagine taking an infinite number of samples from the ***null distribution***



The law of large numbers:



$\bar{X}_n$ □ μ as n →∞ with probability of 1. In words this means that the sample mean converges to the true mean …eventually (with a large enough sample)

The greater the sample size, the greater the precision of the estimate of a proportion. A good explanation of the Law of Large Numbers and the closely related CTL:

https://www.youtube.com/watch?v=9yQpg3z9_DM

Which provides a better strategy for inferential statistics?

Both strategies cost the same, but which one is ’better’ sampling?

Sample 1000 genomes 5 times

or

Sample 5 genomes 1000 times?

Explain your answer. (This simulation doesn’t match to these sample sizes exactly, but it might help to compare the extreme sizes: https://onlinestatbook.com/stat_sim/sampling_dist/index.html)