Module 3C: Thinking in Distributions

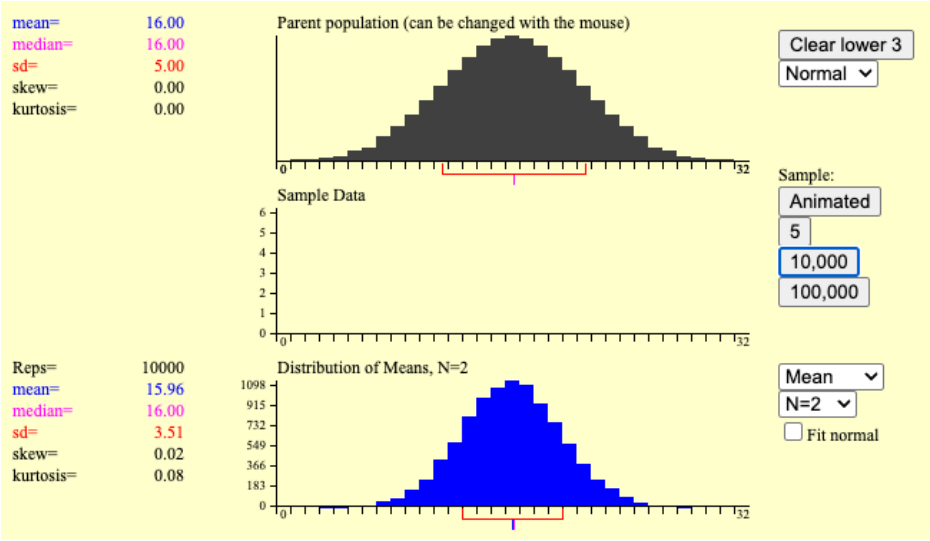
Building block for Hypothesis Testing

Agenda:

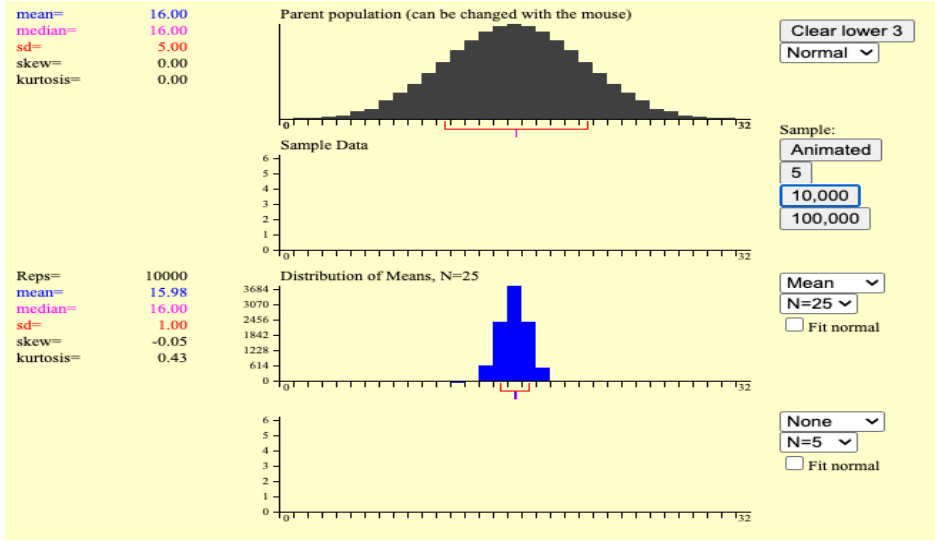
- Major distributions:
  - **Discrete Distributions**
    - Bernoulli
    - Binomial
    - Poisson
    - Hypergeometric
  - **Continuous Distributions**
    - Normal
    - Uniform
    - Exponential
    - Gamma
- Interactive simulations
- **Central Limit Theorem**
  - **Sampling Distribution of the mean**

Central Limit Theorem

- CLT allows us to assume that any sampling distribution of the mean is normally distributed….
  - https://onlinestatbook.com/stat_sim/sampling_dist/



- sampling n=2, 10000 times

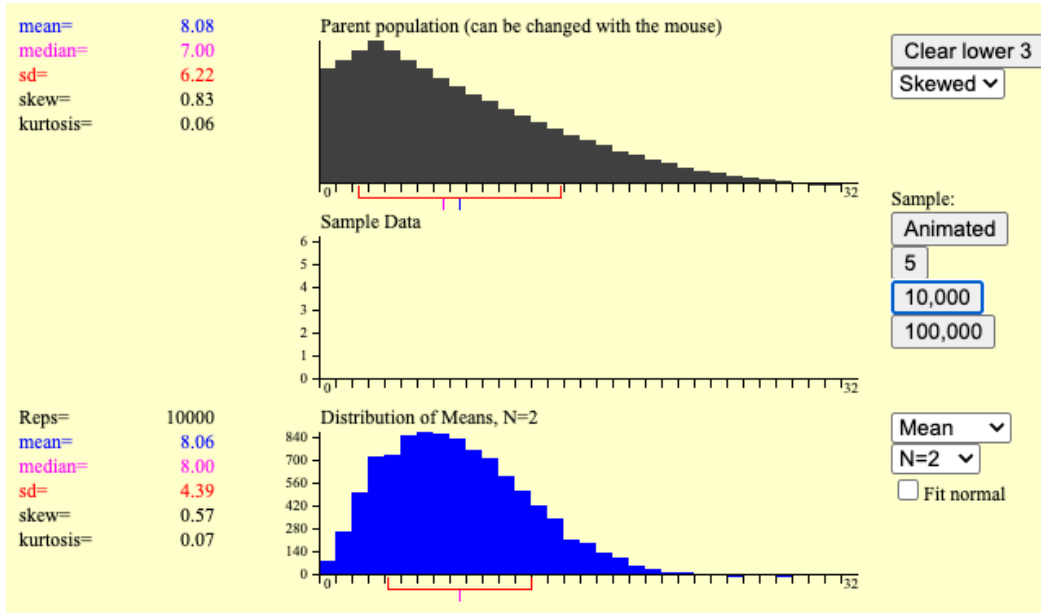


- sampling n=25, 10,000 times

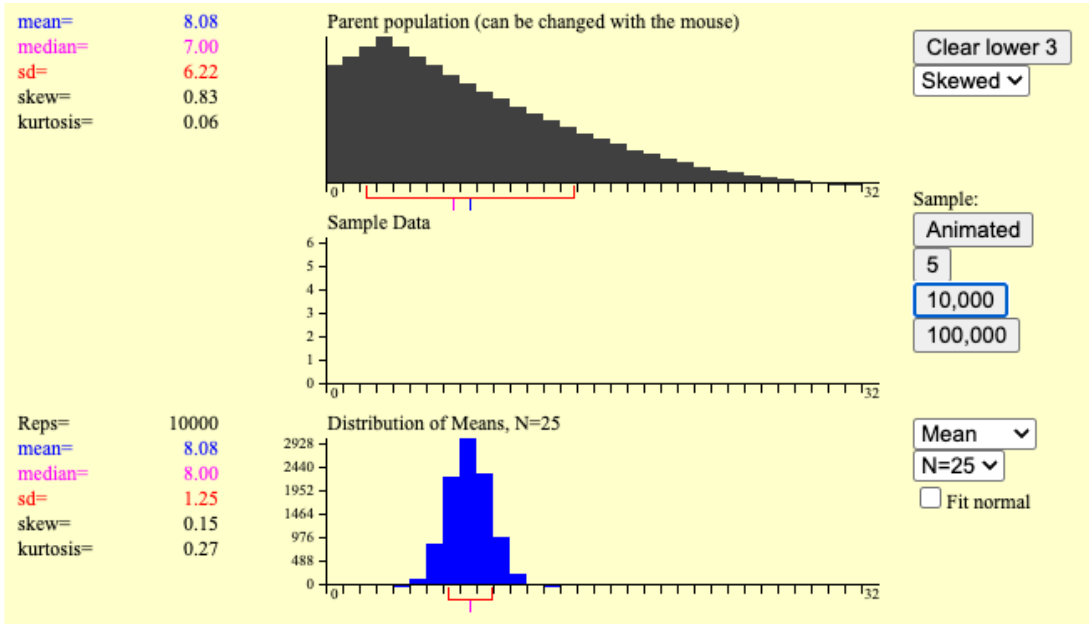
both from Normal Distributed Variable, but have different $SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

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)
  - https://onlinestatbook.com/stat_sim/sampling_dist/



- sampling n=2, 10000 times



- sampling n=25, 10,000 times

both from Skewed distributed Variable, but have different $SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

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)
  - https://onlinestatbook.com/stat_sim/sampling_dist/
- 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.

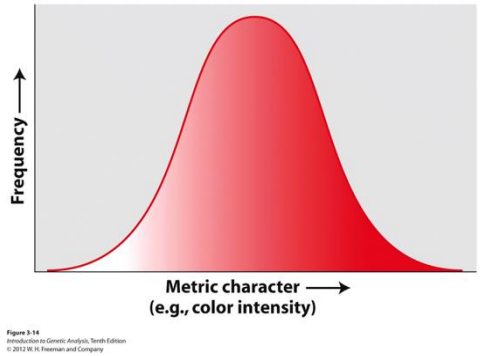
Fisher's paper is challenging to read but, for our interest, we can think about it as a quincunx simulator, a simple demonstration of how to get a normal distribution from multiple discrete alleles.

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

* *You can think of the peak on a quincunx as the result of 'counting paths'.* ***There are more paths (choices) that end up in the center and since each path is additive it results in the summative phenotypic peak and fewer paths that end up with extreme phenotypes at either edge.*** *This is what we do with classic Punnett squares on the next slide.*

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).



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:

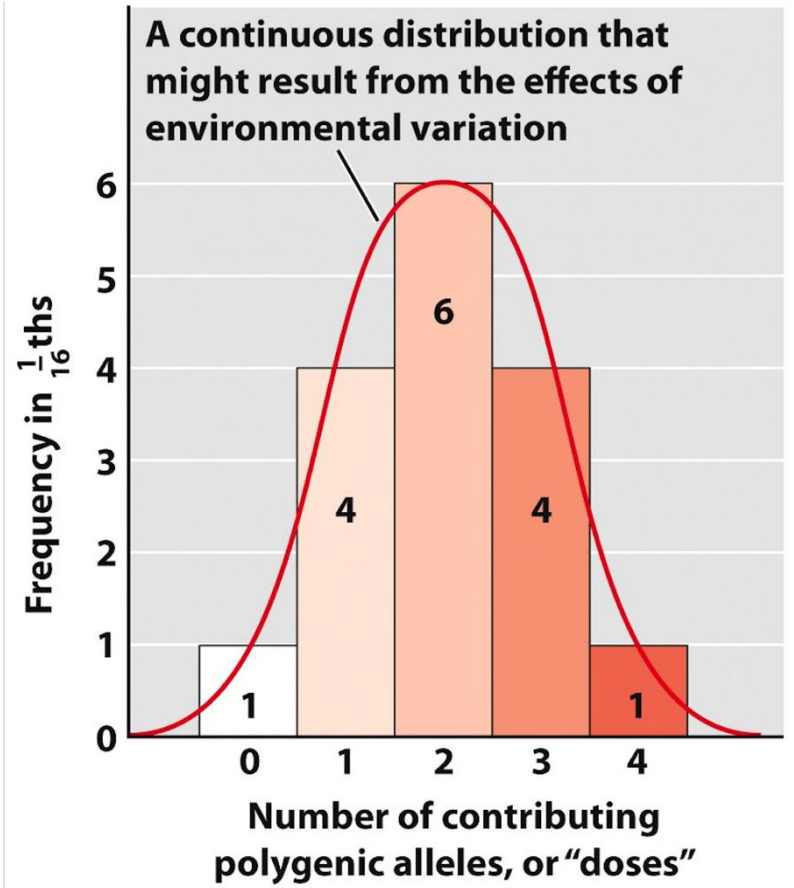
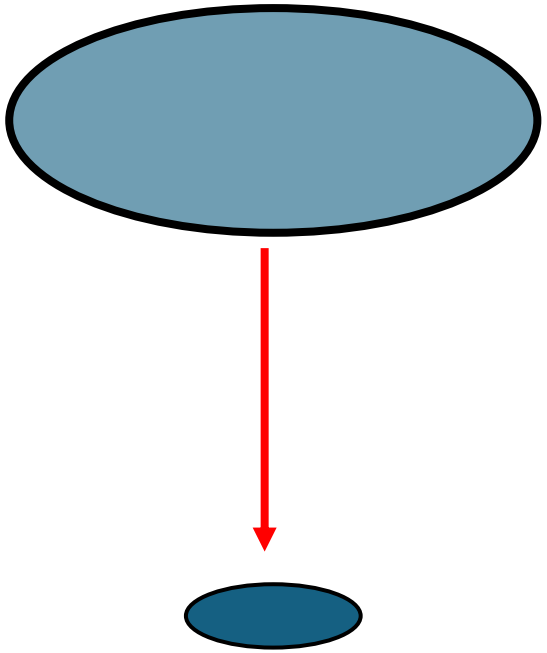


Figure 3-16
Introduction to Genetic Analysis, Tenth Edition

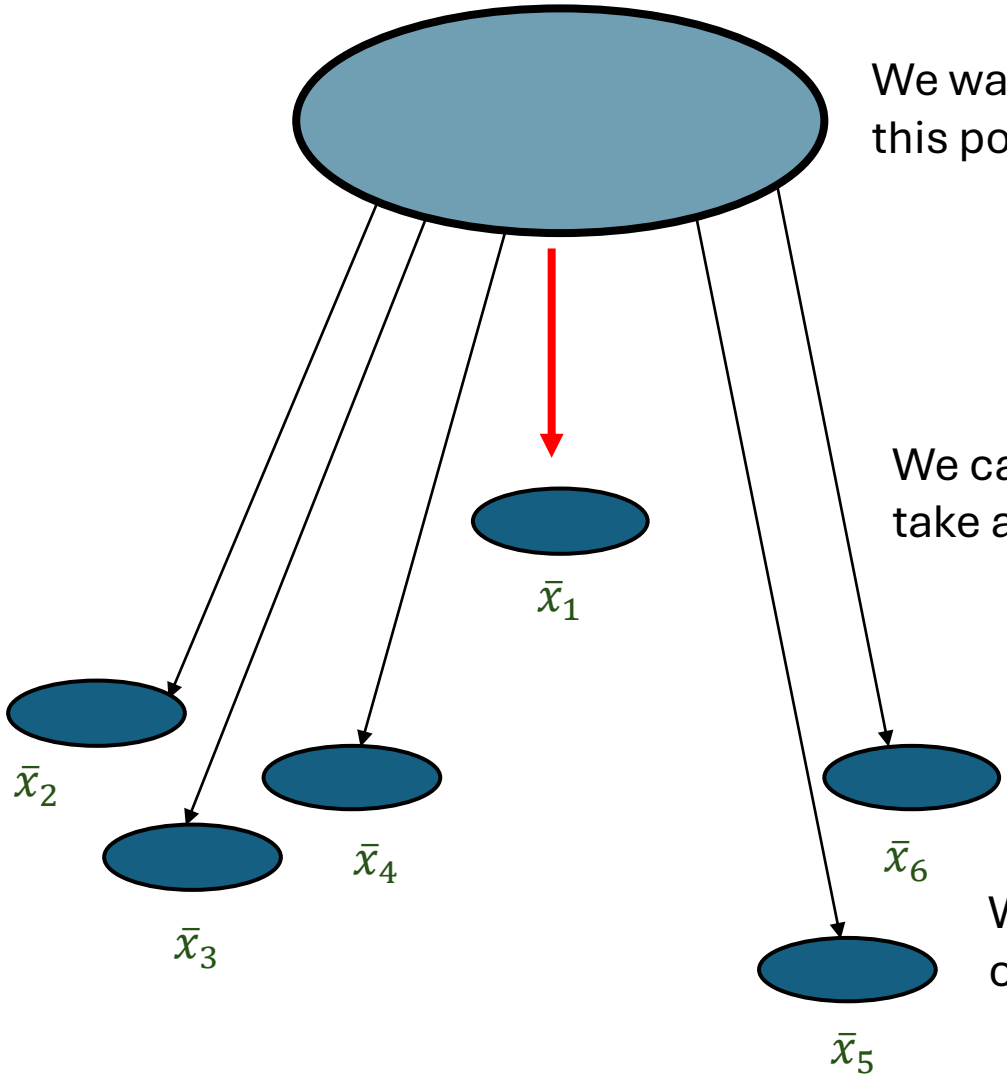
Sampling matters: How good is your slice of reality?



We want to know something about this population, typically μ

We can't measure everyone, so we take a sample, typically $\bar{x}$

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

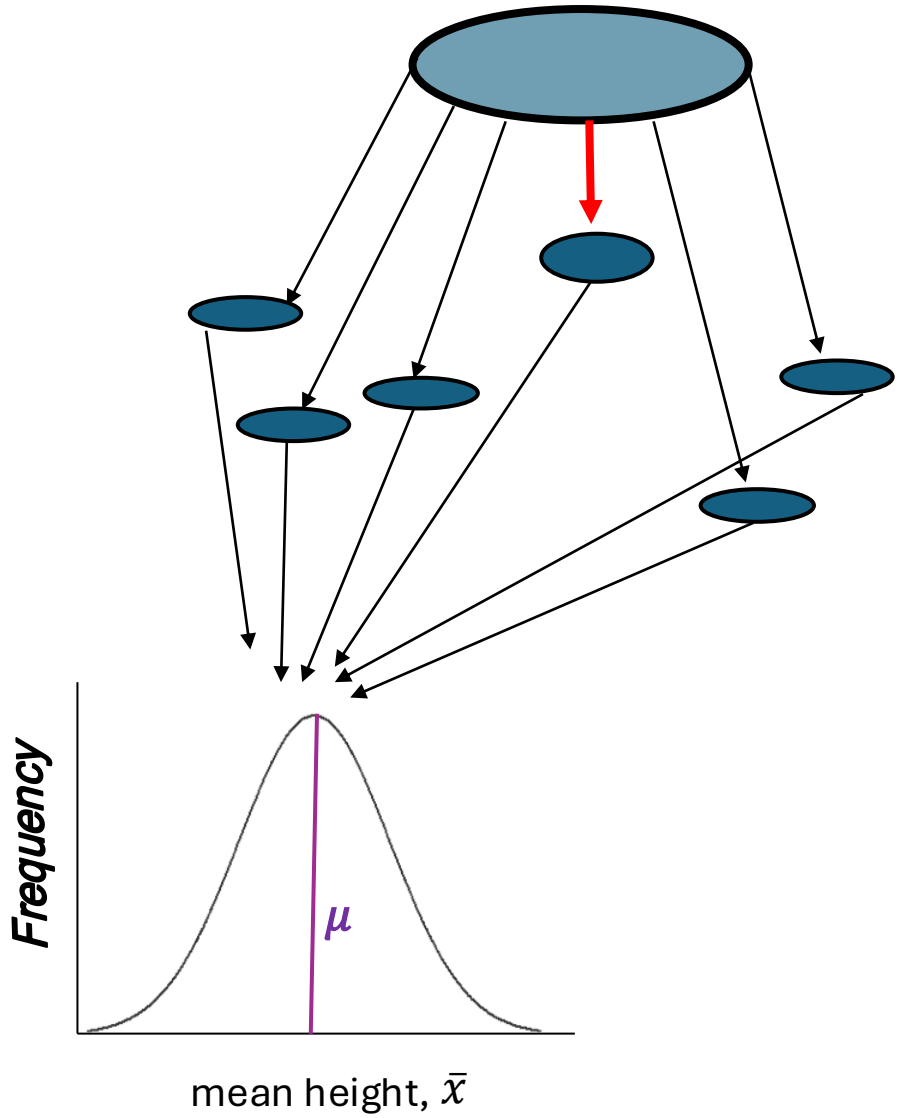


We want to know something about this population, μ

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

How '*good* 'is the sample?

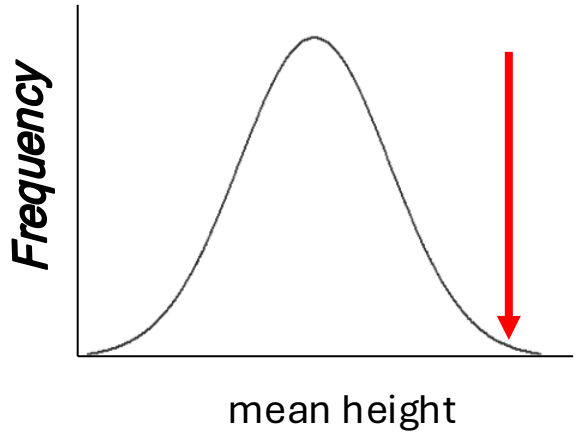
We imagine taking an infinite number of samples from the ***null distribution***



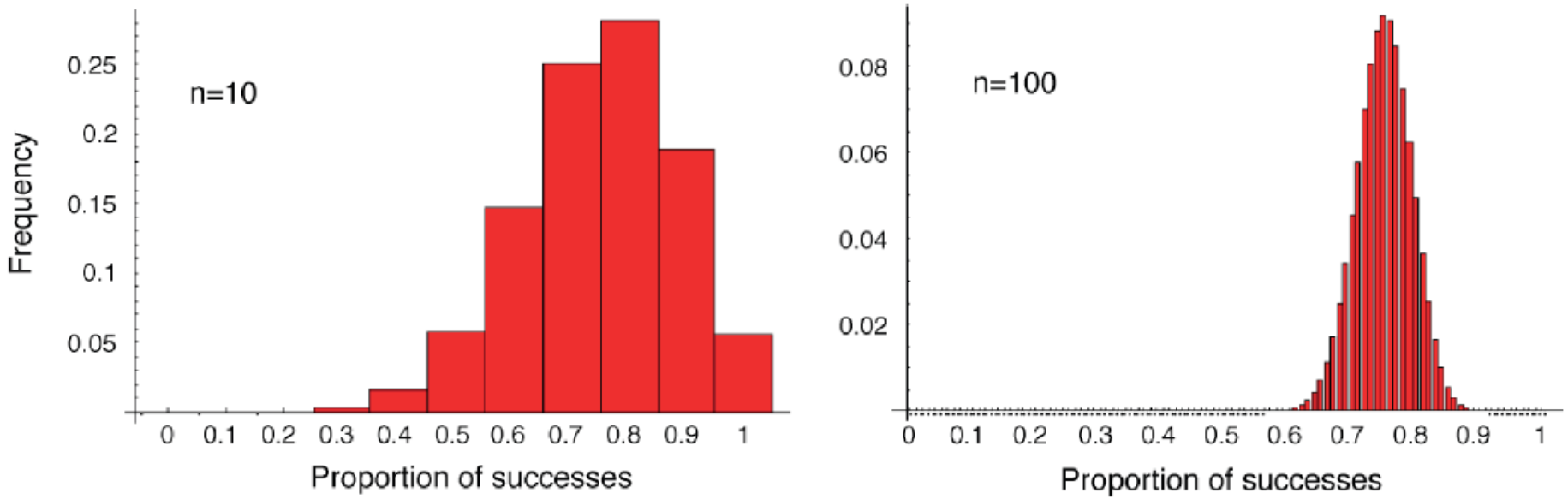
- We want to know something about this population
- We can't measure everyone, so we take a sample

How 'good' is the sample?

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



The law of large numbers:



$\bar{X}_n \to \mu$ 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=VpuN8vCQ--M

https://www.youtube.com/watch?v=YAlJCEDH2uY

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)