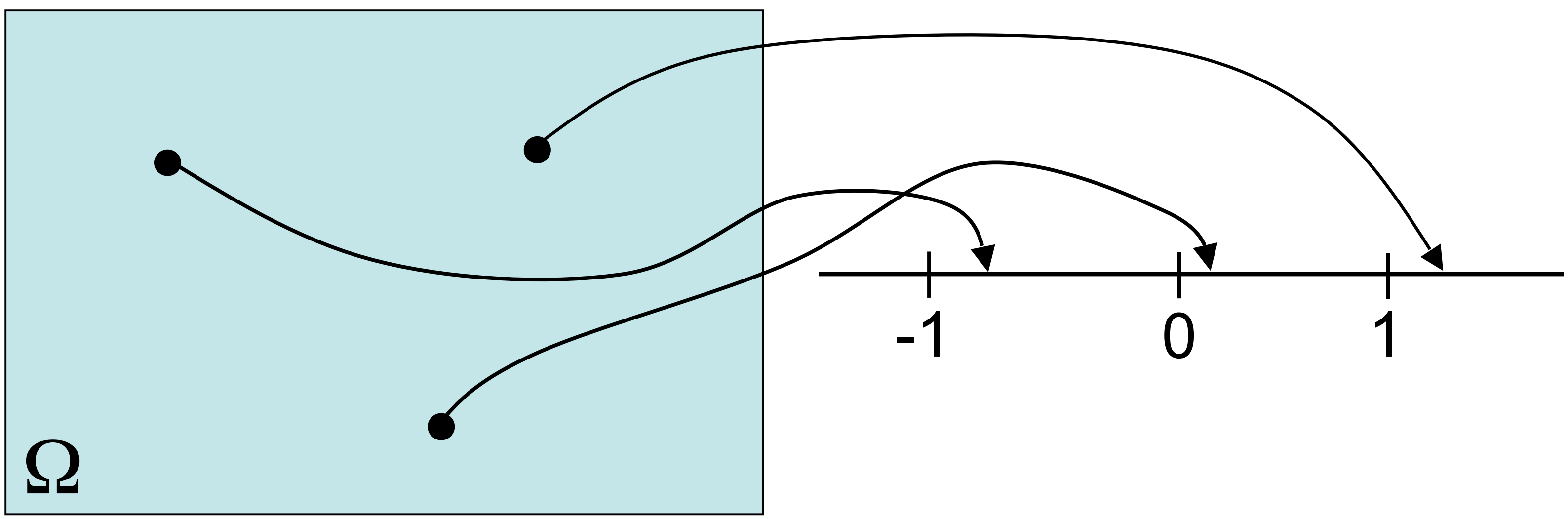
Lecture 4: Random Variables and Distributions

Goals

- Random Variables
- Overview of discrete and continuous distributions important in genetics/genomics
- Working with distributions in R

Random Variables

A rv is any rule (i.e., function) that associates a number with each outcome in the sample space



Two Types of Random Variables

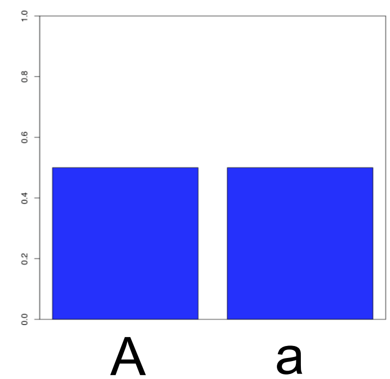
- A ***discrete*** random variable has a ***countable*** number of possible values
- A ***continuous*** random variable takes all values in an interval of numbers

Probability Distributions of RVs

Discrete

Let X be a discrete rv. Then the *probability mass function (pmf),* f(x), of X is:

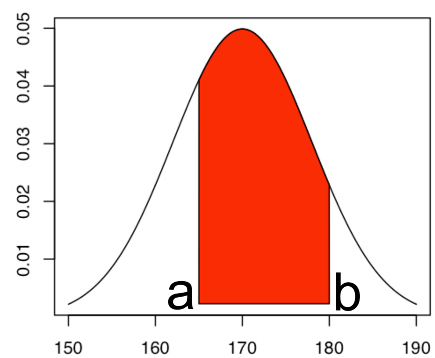
$$f(x) = \begin{cases} P(X = x), & x \in \Omega \\ 0, & x \notin \Omega \end{cases}$$



Continuous

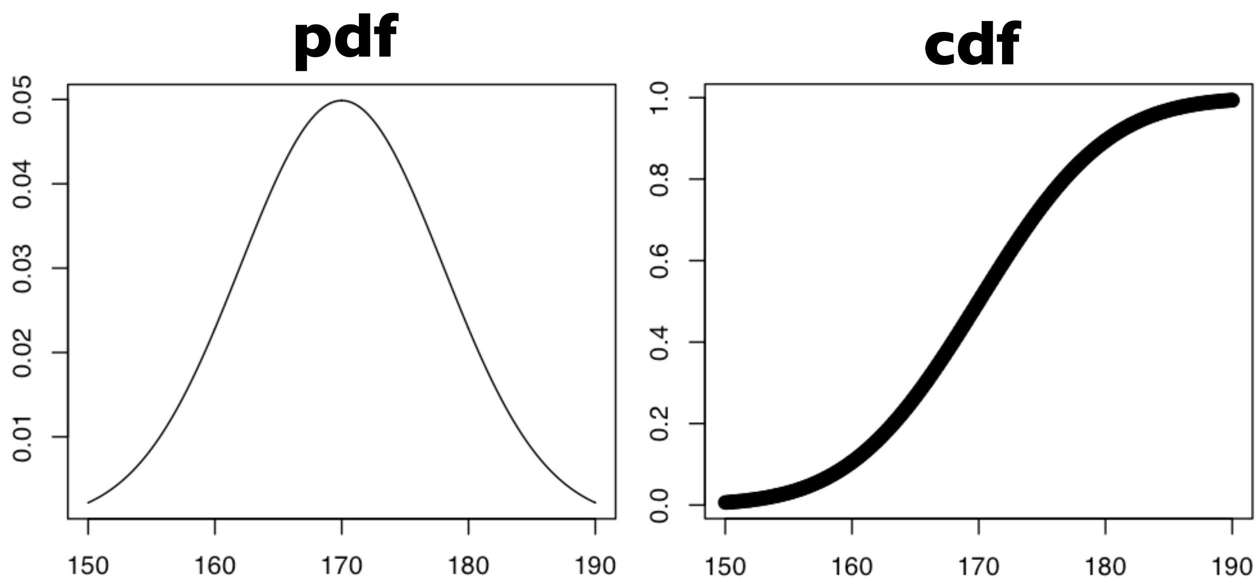
Let X be a continuous rv. Then the *probability density function (pdf)* of X is a function f(x) such that for any two numbers a and b with a ≤ b:

$$P(a \le X \le b) = \int_a^b f(x)dx$$



Using CDFs to Compute Probabilities

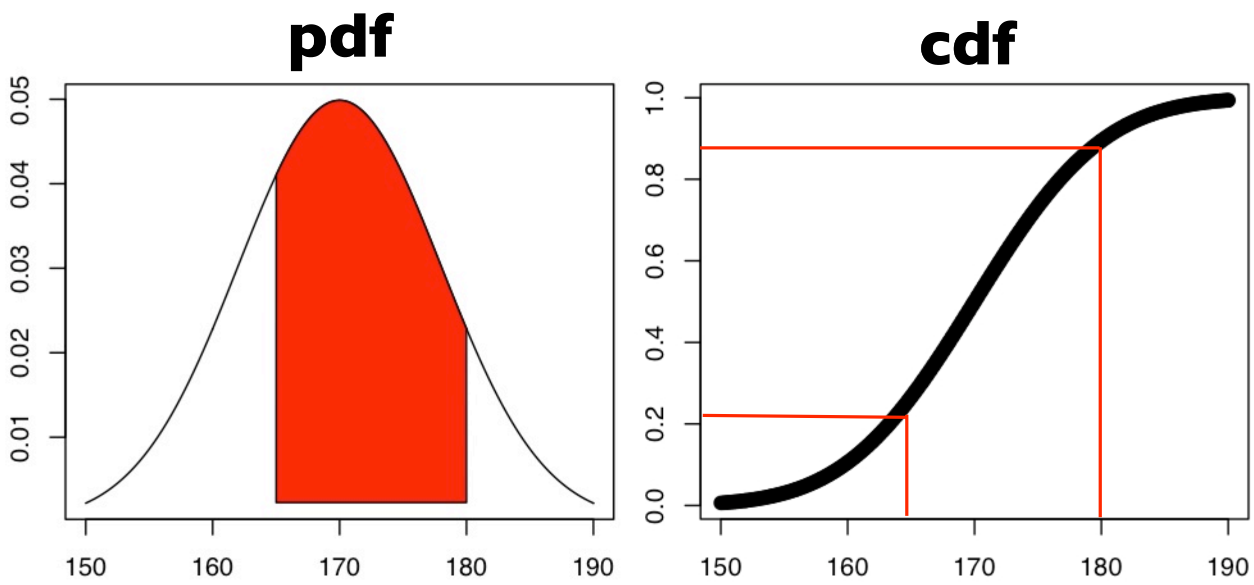
Continuous rv: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$



$$P(a \leq X \leq b) = F(b) - F(a)$$

Using CDFs to Compute Probabilities

Continuous rv: $F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy$



$$P(a \le X \le b) = F(b) - F(a)$$

Expectation of Random Variables

Discrete

Let X be a discrete rv that takes on values in the set D and has a pmf f(x). Then the expected or mean value of X is:

$$\mu_X = E[X] = \sum_{x \in D} x \cdot f(x)$$

Continuous

The expected or mean value of a continuous rv X with pdf f(x) is:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Variance of Random Variables

Discrete

Let X be a discrete rv with pmf f(x) and expected value μ. The variance of X is:

$$\sigma_X^2 = V[X] = \sum_{x \in D} (x - \mu)^2 = E[(X - \mu)^2]$$

Continuous

The variance of a continuous rv X with pdf f(x) and mean μ is:

$$\sigma_X^2 = V[X] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx = E[(X - \mu)^2]$$

Example of Expectation and Variance

- Let $L_1, L_2, \ldots, L_n$ be a sequence of n nucleotides and define the rv X_i:

$$X_i \begin{cases} 1, \text{ if } L_i = A \\ 0, \text{ otherwise} \end{cases}$$

- pmf is then: $P(X_i = 1) = P(L_i = A) = p_A$

$$P(X_i = 0) = P(L_i = C \text{ or } G \text{ or } T) = 1 - p_A$$

- $E[X] = 1 \text{ x } p_A + 0 \text{ x } (1 - p_A) = p_A$

- $Var[X] = E[X - \mu]^2 = E[X^2] - \mu^2$

$$= [1^2 \text{ x } p_A + 0^2 \text{ x } (1 - p_A)] - p_A^2$$

$$= p_A (1 - p_A)$$

The Distributions We'll Study

1. Binomial Distribution

2. Hypergeometric Distribution

3. Poisson Distribution

4. Normal Distribution

Binomial Distribution

- **Experiment consists of n trials**
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- **Trials are identical and each can result in one of the same two outcomes**
  - e.g., head or tail in each toss of a coin
  - Generally called “success” and “failure”
  - Probability of success is p, probability of failure is 1 – p
- **Trials are independent**
- **Constant probability for each observation**
  - e.g., Probability of getting a tail is the same each time we toss the coin

Binomial Distribution

pmf:

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

cdf:

$$P\{X \le x\} = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y}$$

E(x) = np

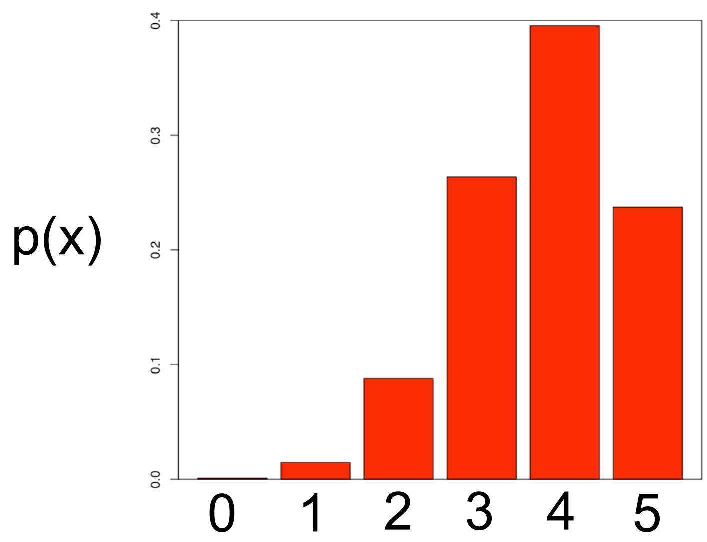
Var(x) = np(1-p)

Binomial Distribution: Example 1

- A couple, who are both carriers for a recessive disease, wish to have 5 children. They want to know the probability that they will have four healthy kids

$$P\{X = 4\} = \binom{5}{4} 0.75^4 \times 0.25^1$$

$$= 0.395$$



Binomial Distribution: Example 2

- Wright-Fisher model: There are i copies of the A allele in a population of size 2N in generation t. What is the distribution of the number of A alleles in generation t + 1?

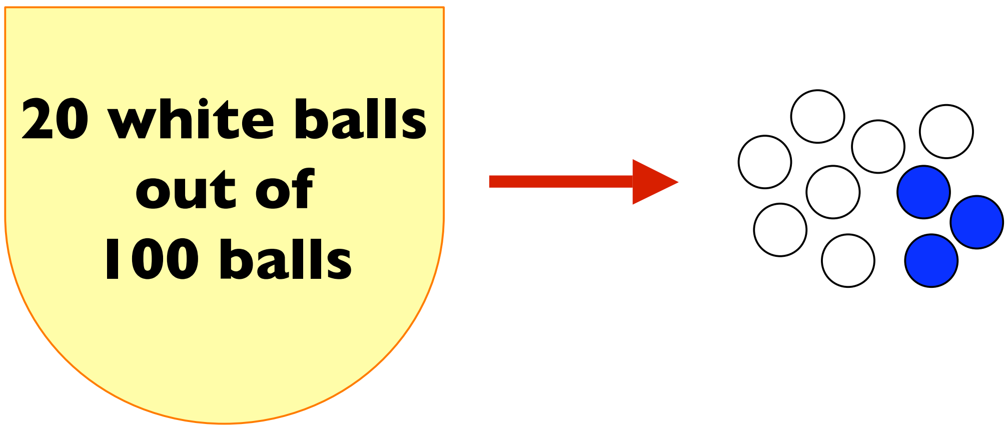
$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^{j} \left(1 - \frac{i}{2N}\right)^{2N-j} \qquad j = 0, 1, \ldots, 2N$$

Hypergeometric Distribution

- **Population to be sampled consists of N finite individuals, objects, or elements**
- **Each individual can be characterized as a success or failure, m successes in the population**
- **A sample of size k is drawn and the rv of interest is X = number of successes**

Hypergeometric Distribution

- Similar in spirit to Binomial distribution, but from a ***finite*** population ***without*** replacement



If we randomly sample 10 balls, what is the probability that 7 or more are white?

Hypergeometric Distribution

- pmf of a hypergeometric rv:

$$P\{X = i \mid n, m, k\} = \frac{\binom{m}{i}\binom{n}{k-i}}{\binom{m+n}{k}}$$

For i = 0, 1, 2, 3, …

Where,

k = Number of balls selected

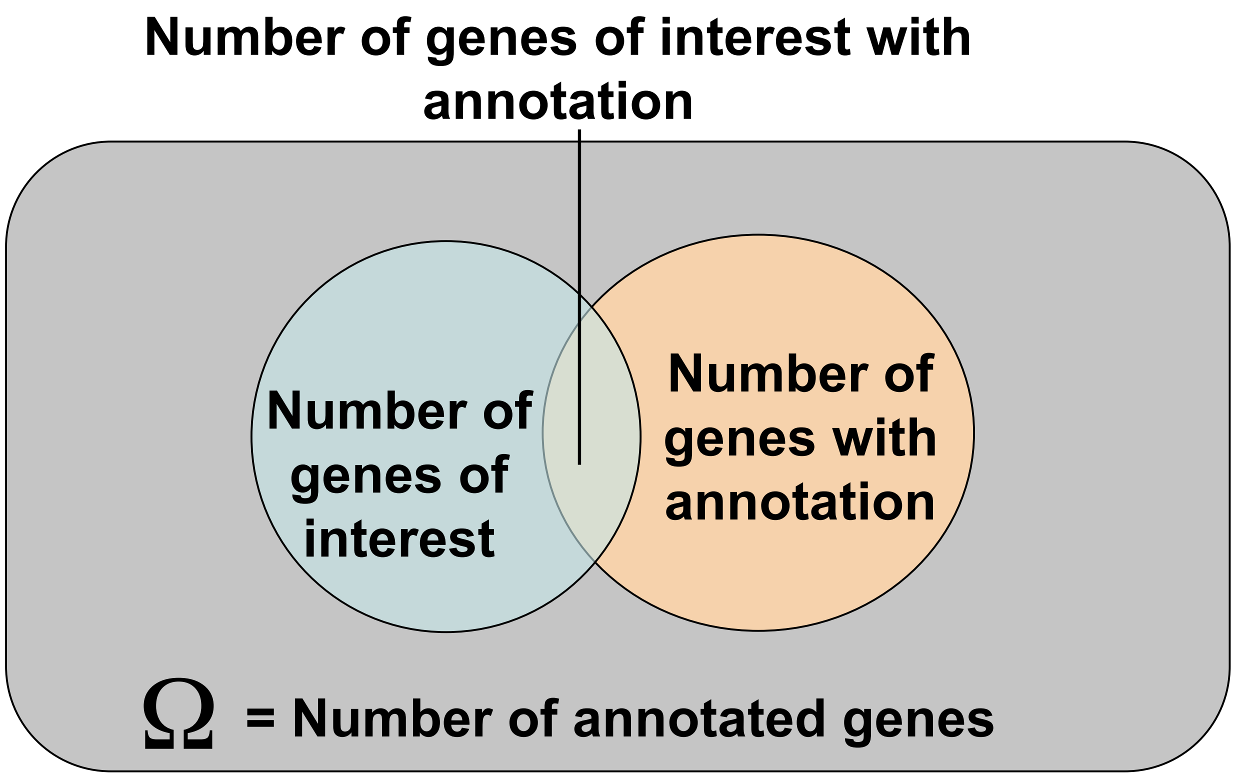
m = Number of balls in urn considered "success"

n = Number of balls in urn considered "failure"

m + n = Total number of balls in urn

Hypergeometric Distribution

- Extensively used in genomics to test for "enrichment":



Poisson Distribution

- **Useful in studying rare events**
- **Poisson distribution also used in situations where "events" happen at certain points in time**
- **Poisson distribution approximates the binomial distribution when n is large and p is small**

Poisson Distribution

- A rv X follows a Poisson distribution if the pmf of X is:

$$P\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!} \qquad \text{For } i = 0, 1, 2, 3, \ldots$$

- λ is frequently a rate per unit time:

  $\lambda = \alpha t$ = expected number of events per unit time t

- Safely approximates a binomial experiment when $n > 100$, $p < 0.01$, $np = \lambda < 20$)

- $E(X) = Var(X) = \lambda$

Poisson RV: Example 1

- The number of crossovers, X, between two markers is X ~ poisson(λ=d)

$$P\{X = i\} = e^{-d} \frac{d^i}{i!}$$

$$P\{X = 0\} = e^{-d}$$

$$P\{X \geq 1\} = 1 - e^{-d}$$

Poisson RV: Example 2

- Recent work in Drosophila suggests the spontaneous rate of deleterious mutations is ~ 1.2 per diploid genome. Thus, let's tentatively assume X ~ poisson(λ = 1.2) for humans. What is the probability that an individual has 12 or more spontaneous deleterious mutations?

$$P\{X \geq 12\} = 1 - \sum_{i=0}^{11} e^{-1.2} \frac{1.2^i}{i!}$$

$$= 6.17 \times 10^{-9}$$

Poisson RV: Example 3

- Suppose that a rare disease has an incidence of 1 in 1000 people per year. Assuming that members of the population are affected independently, find the probability of k cases in a population of 10,000 (followed over 1 year) for k=0,1,2.

The expected value (mean) = λ = .001*10,000 = 10

$$P(X=0)=\frac{(10)^0 e^{-(10)}}{0!}=.0000454$$

$$P(X=1)=\frac{(10)^1 e^{-(10)}}{1!}=.000454$$

$$P(X=2)=\frac{(10)^2 e^{-(10)}}{2!}=.00227$$

Normal Distribution

- **“Most important” probability distribution**
- **Many rv’s are approximately normally distributed**
- **Even when they aren’t, their sums and averages often are (CLT)**

Normal Distribution

- pdf of normal distribution:

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

- standard normal distribution ($\mu = 0$, $\sigma^2 = 1$):

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/2}$$

- cdf of Z:

$$P(Z \leq z) = \int_{-\infty}^{z} f(y;0,1)\,\mathrm{dy}$$

Standardizing Normal RV

- If X has a normal distribution with mean μ and standard deviation σ, we can standardize to a standard normal rv:

$$Z = \frac{X - \mu}{\sigma}$$

I Digress: Sampling Distributions

- Before data is collected, we regard observations as random variables ($X_1, X_2, \ldots, X_n$)

- This implies that until data is collected, any function (statistic) of the observations (mean, sd, etc.) is also a random variable

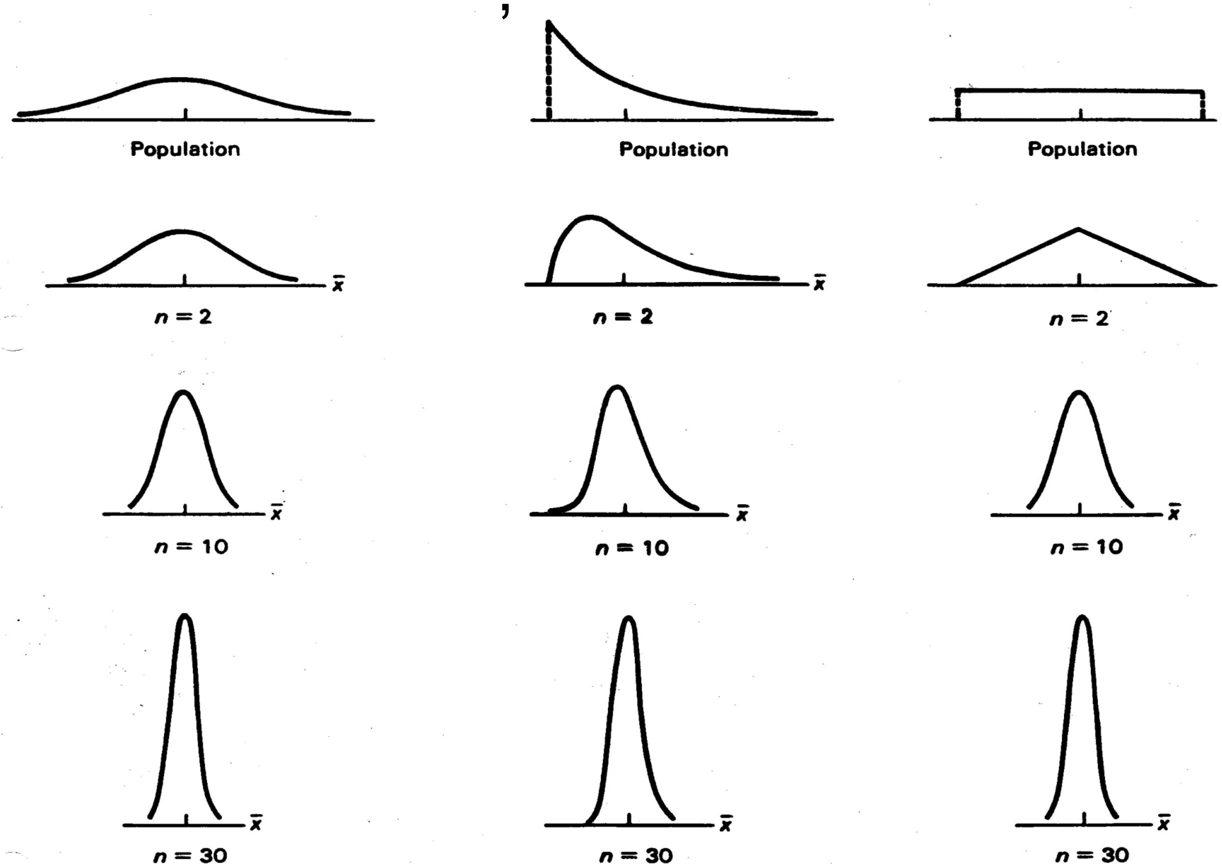
- Thus, any statistic, because it is a random variable, has a probability distribution - referred to as a **sampling distribution**

- Let's focus on the sampling distribution of the mean, $\bar{X}$

Behold The Power of the CLT

- Let $X_1, X_2, \ldots, X_n$ be an iid random sample from a distribution with mean μ and standard deviation σ. If n is sufficiently large:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



Example

- If the mean and standard deviation of serum iron values from healthy men are 120 and 15 mgs per 100ml, respectively, what is the probability that a random sample of 50 normal men will yield a mean between 115 and 125 mgs per 100ml?

First, calculate mean and sd to normalize (120 and $15/\sqrt{50}$)

$$p(115 \le \bar{x} \le 125 = p\left(\frac{115-120}{2.12} \le \bar{x} \le \frac{125-120}{2.12}\right)$$

$$= p(-2.36 \le z \le 2.36)$$

$$= p(z \le 2.36) - p(z \le -2.36)$$

$$= 0.9909 - 0.0091$$

$$= 0.9818$$

R

- **Understand how to calculate probabilities from probability distributions**
  - Normal: dnorm and pnorm
  - Poisson: dpois and ppois
  - Binomial: dbinom and pbinom
  - Hypergeometric: dhyper and phyper

- **Exploring relationships among distributions**