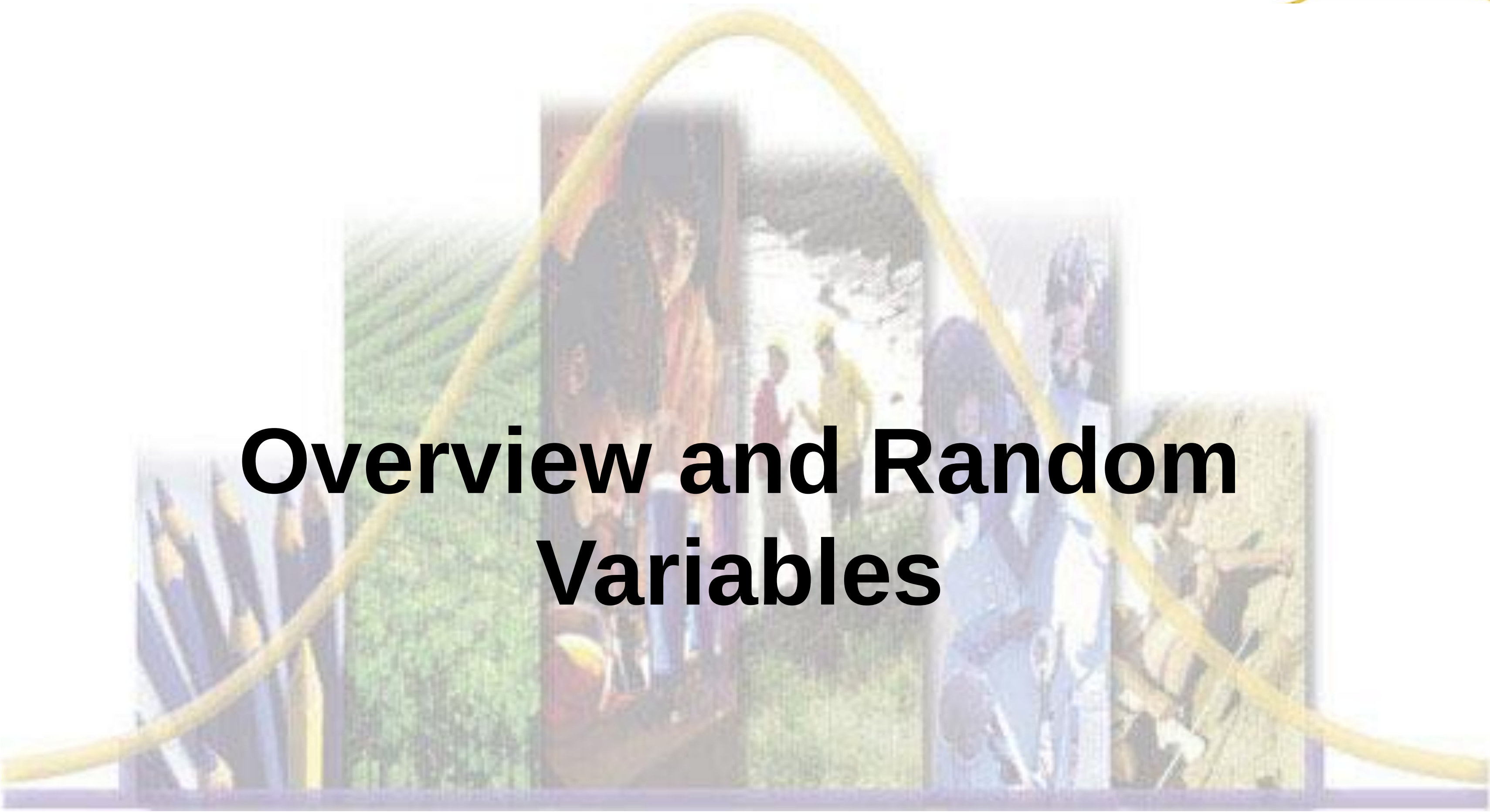
Lecture 4:
Probabilistic features of Distributions

Probability Distributions



Overview and Random Variables

Overview

- Key words

  Probability distribution , random variable ,

  Bernolli distribution, Binomail distribution,

  Poisson distribution

Probability Distributions will describe what will *probably* happen instead of what actually *did* happen.

Definitions

- **A random variable is a variable (typically represented by *X*) that has a single numerical value, determined by chance, for each outcome of a procedure.**

- **A probability distribution is a graph, table, or formula that gives the probability for each value of the random variable.**

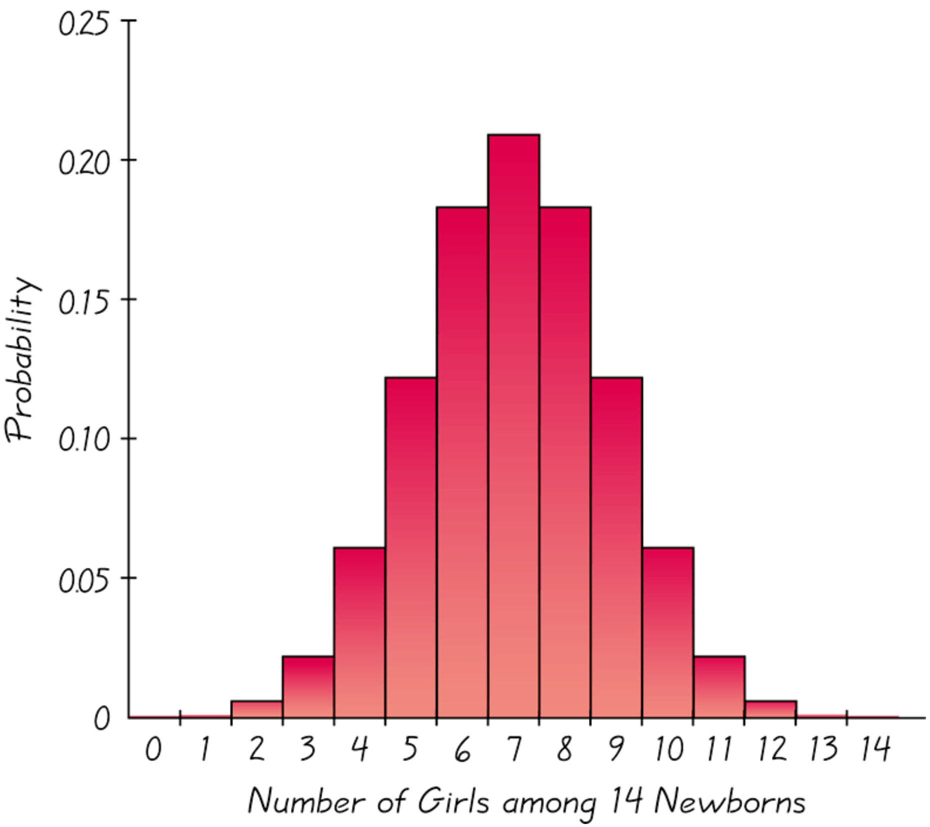
Definitions

- A **discrete random variable** has either a finite number of values or countable number of values, where "countable" refers to the fact that there might be infinitely many values, but they result from a counting process.

- A **continuous random variable** has infinitely many values, and those values can be associated with measurements on a continuous scale in such a way that there are no gaps or interruptions.

Graphs

The **probability histogram** is very similar to a relative frequency histogram, but the vertical scale shows probabilities.



Requirements for Probability Distribution

$$\sum P(x) = 1$$

where x assumes all possible values

$$0 \leq P(x) \leq 1$$

for every individual value of x

Mean, Variance and Standard Deviation of a Probability Distribution

$$\mu = \Sigma\, [x \cdot P(x)]$$ **Mean**

$$\sigma^2 = \Sigma\, [(x - \mu)^2 \cdot P(x)]$$ **Variance**

$$\sigma^2 = [\Sigma\, x^2 \cdot P(x)] - \mu^2$$ **Variance (shortcut)**

$$\sigma = \sqrt{\Sigma\, [x^2 \cdot P(x)] - \mu^2}$$ **Standard Deviation**

Cumulative Probability Distribution of X F(x)

- It shows the probability that the variable X is less than or equal to a certain value, $P(X \le x)$.

Example

Number of Programs	**frequency**	**P(X=x)**	**F(x)= P(X≤ x)**
1	62	0.2088	0.2088
2	47	0.1582	0.3670
3	39	0.1313	0.4983
4	39	0.1313	0.6296
5	58	0.1953	0.8249
6	37	0.1246	0.9495
7	4	0.0135	0.9630
8	11	0.0370	1.0000
Total	297	1.0000	

- **Properties of probability distribution of discrete random variable.**

1. $0 \le P(X = x) \le 1$
2. $\sum P(X = x) = 1$
3. **P(a ≤ X ≤ b) = P(X ≤ b) – P(X ≤ a-1)**
4. **P(X < b) = P(X ≤ b-1)**

Identifying Unusual Results Range Rule of Thumb

According to the range rule of thumb, most values should lie within 2 standard deviations of the mean.

We can therefore identify "unusual" values by determining if they lie outside these limits:

Maximum usual value = $\mu + 2\sigma$

Minimum usual value = $\mu - 2\sigma$

Identifying Unusual Results Probabilities

Rare Event Rule

If, under a given assumption (such as the assumption that boys and girls are equally likely), the probability of a particular observed event (such as 13 girls in 14 births) is extremely small, we conclude that the assumption is probably not correct.

- **Unusually high: *x* successes among *n* trials is an *unusually high* number of successes if *P*(*x* or more) is very small (such as 0.05 or less).**

- **Unusually low: *x* successes among *n* trials is an *unusually low* number of successes if *P*(*x* or fewer) is very small (such as 0.05 or less).**

Definition

The **expected value** of a discrete random variable is denoted by E, and it represents the average value of the outcomes. It is obtained by finding the value of $\Sigma\,[x \cdot P(x)]$.

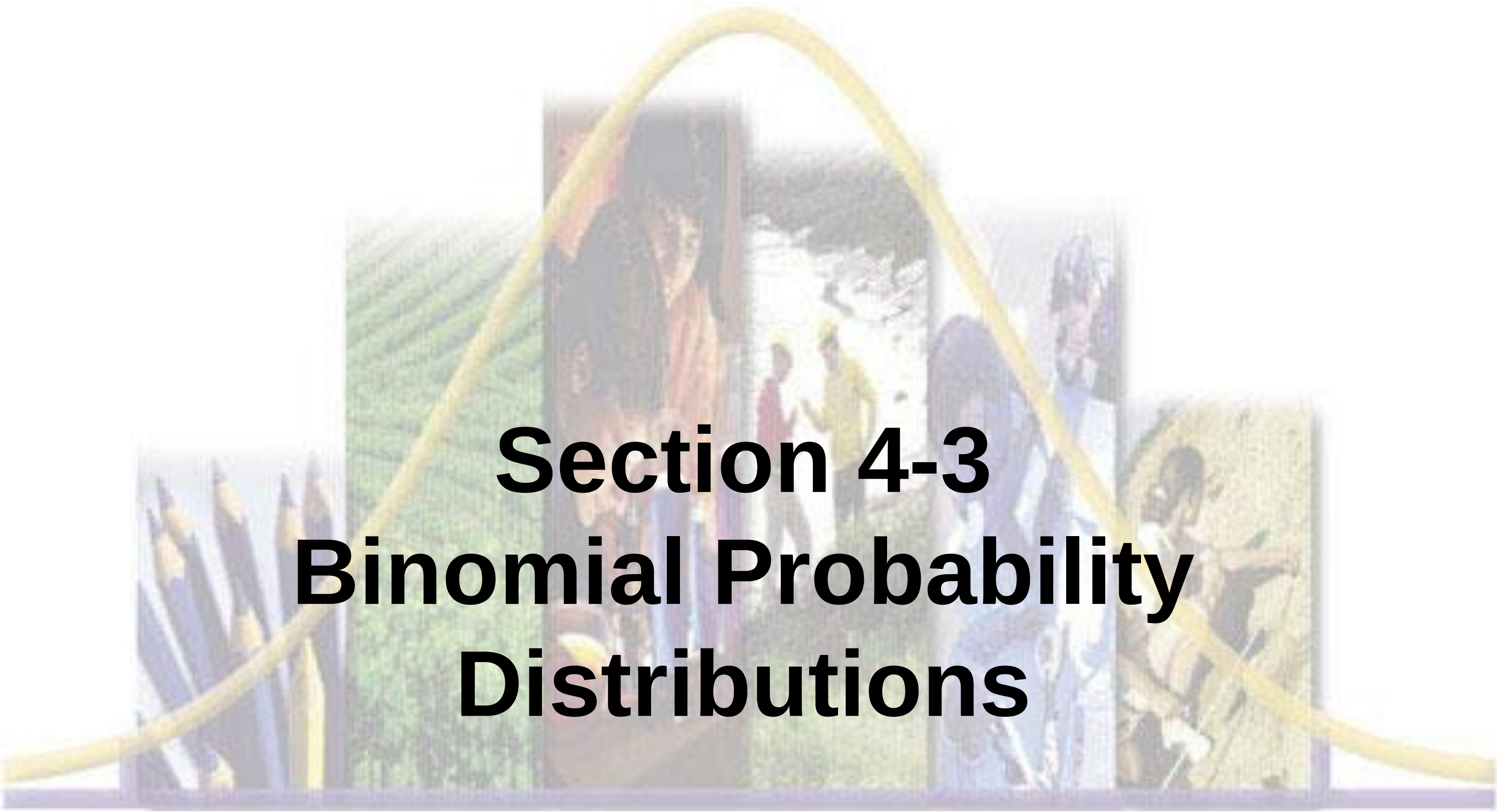
$$E\,(x) = \Sigma\,[x \cdot P(x)]$$

Probability density function

- If *X* is a continuous random variable, then it has a probability density function $f(x)$, and therefore its probability of falling into a given interval, say [*a*, *b*] is given by the integral

$$\Pr[a \le X \le b] = \int_a^b f(x)\,dx$$

- In particular, the probability for *X* to take any single value *a* (that is $a \le X \le a$) is zero, because an integral with coinciding upper and lower limits is always equal to zero.



Section 4-3
Binomial Probability Distributions

Definitions

A binomial probability distribution results from a procedure that meets all the following requirements:

1. **The procedure has a *fixed number of trials*.**
2. **The trials must be *independent*. (The outcome of any individual trial doesn't affect the probabilities in the other trials.)**
3. **Each trial must have all outcomes classified into *two categories*.**
4. **The probabilities must remain *constant* for each trial.**

When a random process or experiment called a trial can result in only one of two mutually exclusive outcomes, such as dead or alive, sick or well, the trial is called a Bernoulli trial.

The Bernoulli Process

- A sequence of Bernoulli trials forms a Bernoulli process under the following conditions

1- Each trial results in one of two possible, mutually exclusive, outcomes. One of the possible outcomes is denoted (arbitrarily) as a success, and the other is denoted a failure.

2- The probability of a success, denoted by p, remains constant from trial to trial. The probability of a failure, 1-p, is denoted by q.

3- The trials are independent, that is the outcome of any particular trial is not affected by the outcome of any other trial

Notation for Binomial Probability Distributions

S and F (success and failure) denote two possible categories of all outcomes; *p* and *q* will denote the probabilities of S and F, respectively, so

$P(\text{S}) = p$ (p = probability of success)

$P(\text{F}) = 1 - p = q$ (q = probability of failure)

Notation (cont)

n **denotes the number of fixed trials.**

x **denotes a specific number of successes in *n* trials, so *x* can be any whole number between 0 and *n*, inclusive.**

p **denotes the probability of *success* in *one* of the *n* trials.**

q **denotes the probability of *failure* in *one* of the *n* trials.**

***P*(*x*)** **denotes the probability of getting exactly *x* successes among the *n* trials.**

Important Hints

- Be sure that *x* and *p* both refer to the same category being called a success.

- When sampling without replacement, the events can be treated as if they were independent if the sample size is no more than 5% of the population size. (That is *n* is less than or equal to 0.05*N*.)

Methods for Finding Probabilities

We will now present two methods for finding the probabilities corresponding to the random variable *x* in a binomial distribution.

Method 1: Using the Binomial Probability Formula

$$P(x) = \frac{n!}{(n-x)!x!} \cdot p^x \cdot q^{n-x}$$

for $x = 0, 1, 2, \ldots, n$

where

n = number of trials

x = number of successes among n trials

p = probability of success in any one trial

q = probability of failure in any one trial ($q = 1 - p$)

Method 2: Using Table

Part of Table A-1 is shown below. With *n* = 4 and *p* = 0.2 in the binomial distribution, the probabilities of 0, 1, 2, 3, and 4 successes are 0.410, 0.410, 0.154, 0.026, and 0.002 respectively.

From Table A-1:

n	x	p 0.20
4	0	0.410
	1	0.410
	2	0.154
	3	0.026
	4	0.002

Binomial probability distribution for $n = 4$ and $p = 0.2$

x	$P(x)$
0	0.410
1	0.410
2	0.154
3	0.026
4	0.002

Rationale for the Binomial Probability Formula

$$P(x) = \underbrace{\frac{n!}{(n-x)!x!}}_{\text{Number of outcomes with exactly } x \text{ successes among } n \text{ trials}} \cdot p^x \cdot q^{n-x}$$

Number of outcomes with exactly *x* successes among *n* trials

Binomial Probability Formula

$$P(x) = \underbrace{\frac{n\,!}{(n-x)!\,x!}}_{\substack{\text{Number of}\\ \text{outcomes with}\\ \text{exactly } x\\ \text{successes}\\ \text{among } n \text{ trials}}} \cdot \underbrace{p^{x} \cdot q^{n-x}}_{\substack{\text{Probability of } x\\ \text{successes}\\ \text{among } n \text{ trials}\\ \text{for any one}\\ \text{particular order}}}$$

- The probability distribution of the binomial random variable **X**, the number of successes in **n** independent trials is :

$$f(x) = P(X = x) = \binom{n}{x} p^X q^{n-X} \quad , \quad x = 0,1,2,\dots,n$$

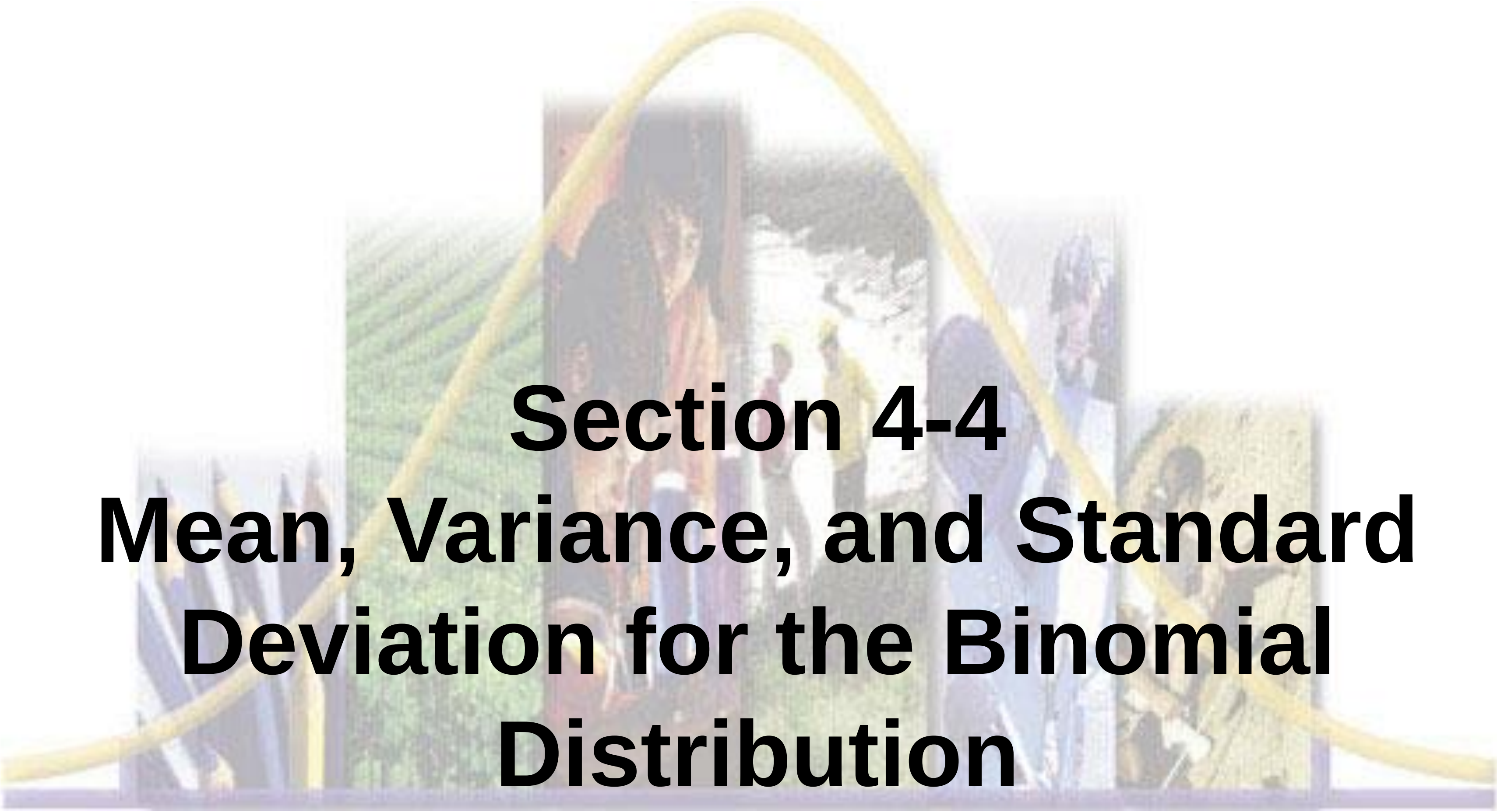
Where $\binom{n}{x}$ is the number of combinations of **n** distinct objects taken **x** of them at **a** time.

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

* Note: 0! =1 $\quad x! = x(x-1)(x-2)\dots(1)$

Properties of the binomial distribution

- 1. $f(x) \geq 0$
- 2. $\sum f(x) = 1$
- 3.The parameters of the binomial distribution are *n* and *p*
- 4. $\mu = E(X) = np$
- 5. $\sigma^2 = \mathrm{var}(X) = np(1-p)$



Section 4-4
Mean, Variance, and Standard Deviation for the Binomial Distribution

Any Discrete Probability Distribution: Formulas

Mean $\mu = \Sigma[x \cdot P(x)]$

Variance $\sigma^2 = [\Sigma x^2 \cdot P(x)] - \mu^2$

Std. Dev $\sigma = \sqrt{[\Sigma x^2 \cdot P(x)] - \mu^2}$

Binomial Distribution: Formulas

Mean $\mu = n \cdot p$

Variance $\sigma^2 = n \cdot p \cdot q$

Std. Dev. $\sigma = \sqrt{n \cdot p \cdot q}$

Where

n **= number of fixed trials**

p **= probability of *success* in one of the *n* trials**

q **= probability of *failure* in one of the *n* trials**

Interpretation of Results

It is especially important to interpret results. The range rule of thumb suggests that values are unusual if they lie outside of these limits:

Maximum usual values $= \mu + 2\,\sigma$

Minimum usual values $= \mu - 2\,\sigma$

Example

Determine whether 68 girls among 100 babies could easily occur by chance.

For this binomial distribution,

μ = 50 girls

σ = 5 girls

$\mu + 2\sigma = 50 + 2(5) = 60$

$\mu - 2\sigma = 50 - 2(5) = 40$

The usual number girls among 100 births would be from 40 to 60. So 68 girls in 100 births is an unusual result.

Limiting distributions

- *de Moivre–Laplace theorem*: As n approaches ∞ while p remains fixed, the distribution of

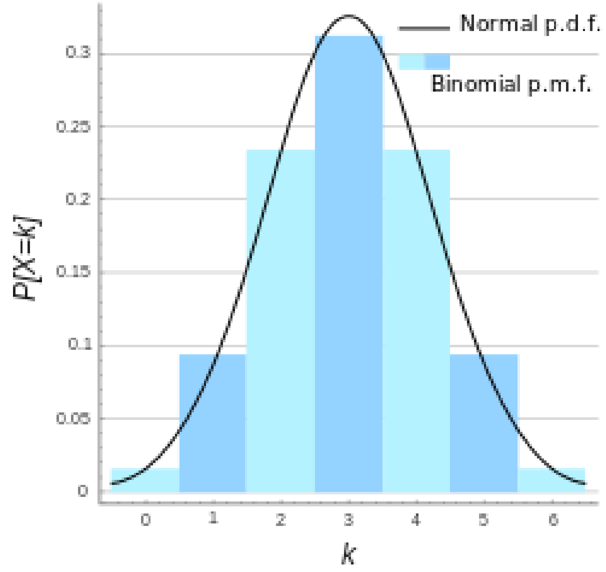
approaches the normal distribution with expected value 0 and variance 1. This result is sometimes loosely stated by saying that the distribution of X is asymptotically normal with expected value np and variance $np(1-p)$. This result is a specific case of the central limit theorem.

Microarray gene expression → normal distribution due to p fixed

Normal approximation

- If n is large enough, then the skew of the distribution is not too great. In this case a reasonable approximation to B(n, p) is given by the normal distribution

$$\mathcal{N}(np,\, np(1-p)),$$



Limiting distributions

- *Poisson limit theorem*: As n approaches ∞ and p approaches 0 while np remains fixed at $\lambda > 0$ or at least np approaches $\lambda > 0$, then the Binomial(n, p) distribution approaches the Poisson distribution with expected value λ.

RNA-seq → Poisson distribution due to average counts fixed

Poisson approximation

- The binomial distribution converges towards the Poisson distribution as the number of trials goes to infinity while the product np remains fixed. Therefore the Poisson distribution with parameter $\lambda = np$ can be used as an approximation to $B(n, p)$ of the binomial distribution if n is sufficiently large and p is sufficiently small. According to two rules of thumb, this approximation is good if $n \geq 20$ and $p \leq 0.05$, or if $n \geq 100$ and $np \leq 10$.

Other distributions

- Hypergeometric distribution
- Negative binomial distribution

Hypergeometric distribution

- The **hypergeometric distribution** is a discrete probability distribution that describes the probability of k successes in draws *without* replacement from a finite population of size N containing exactly K successes.
- This is in contrast to the binomial distribution, which describes the probability of k successes in n draws *with* replacement.

Hypergeometric distribution

- A random variable X follows the hypergeometric distribution :

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

N is the population size
K is the number of success states in the population
n is the number of draws
k is the number of successes

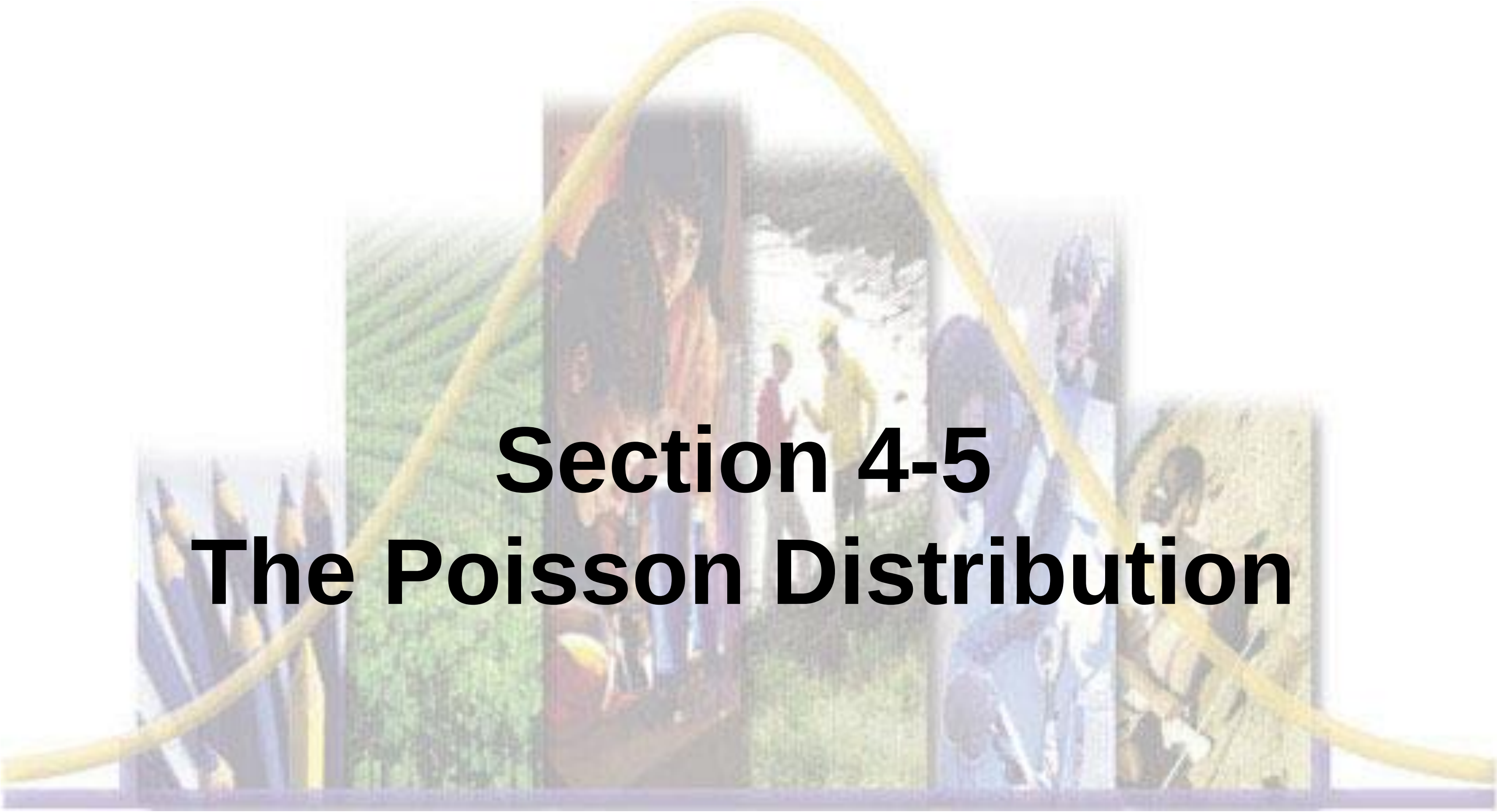
Negative binomial distribution

- **negative binomial distribution** is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures (denoted *r*) occurs. For example, if we define a "1" as failure, and all non "1"s as successes. and we throw a die repeatedly until the third time “1” appears (*r* = three failures), then the probability distribution of the number of non-“1”s that had appeared will be negative binomial.

Negative binomial distribution

- p is the probability of success, and (1-p) is the probability of failure.

$$f(k; r, p) \equiv \Pr(X = k) = \binom{k+r-1}{k} p^k (1-p)^r \quad \text{for } k = 0, 1, 2, \ldots$$



Section 4-5
The Poisson Distribution

Definition

The Poisson distribution is a discrete probability distribution that applies to occurrences of some event *over a specified interval*. The random variable *x* is the number of occurrences of the event in an interval. The interval can be time, distance, area, volume, or some similar unit.

Formula

$$P(x) = \frac{\mu^{x} \cdot e^{-\mu}}{x!} \quad \text{where } e \approx 2.71828$$

Poisson Distribution Requirements

- **The random variable *x* is the number of occurrences of an event *over some interval*.**
- **The occurrences must be *random*.**
- **The occurrences must be *independent* of each other.**
- **The occurrences must be *uniformly distributed* over the interval being used.**

Parameters

- **The mean is μ.**
- **The standard deviation is $\sigma = \sqrt{\mu}$.**

Difference from a Binomial Distribution

The Poisson distribution differs from the binomial distribution in these fundamental ways:

- **The binomial distribution is affected by the sample size *n* and the probability *p*, whereas the Poisson distribution is affected only by the mean *μ*.**
- **In a binomial distribution the possible values of the random variable *x* are 0, 1, . . . *n*, but a Poisson distribution has possible *x* values of 0, 1, . . . , with no upper limit.**

Example

World War II Bombs **In analyzing hits by V-1 buzz bombs in World War II, South London was subdivided into 576 regions, each with an area of 0.25 km^2. A total of 535 bombs hit the combined area of 576 regions**

If a region is randomly selected, find the probability that it was hit exactly twice.

The Poisson distribution applies because we are dealing with occurrences of an event (bomb hits) over some interval (a region with area of 0.25 km^2).

Example

The mean number of hits per region is

$$\mu = \frac{\text{number of bomb hits}}{\text{number of regions}} = \frac{535}{576} = 0.929$$

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} = \frac{0.929^2 \cdot 2.71828^{-0.929}}{2!} = \frac{0.863 \cdot 0.395}{2} = 0.170$$

The probability of a particular region being hit exactly twice is $P(2) = 0.170$.

Poisson as Approximation to Binomial

The Poisson distribution is sometimes used to approximate the binomial distribution when n is large and p is small.

Rule of Thumb

- $n \geq 100$
- $np \leq 10$

Poisson as Approximation to Binomial - μ

- $n \geq 100$
- $np \leq 10$

Value for μ

$$\mu = n \cdot p$$

Properties of the Poisson distribution p(x)=f(x)

- 1. $f(x) \geq 0$
- 2. $\sum f(x) = 1$
- 3. $\mu = E(X) = \lambda$
- 4. $\sigma^2 = \mathrm{var}(X) = \lambda$