Discrete Probability Distributions
Discrete Probability Distribution

• Are used to model outcomes that only have a finite number of possible values.
• For example, the number of congenitally missing third molars in 20 year olds.
• There are only 5 possible values (0, 1, 2, 3, 4).
• The discrete probability distribution assigns a probability to each possible value.
Example: congenitally missing third molars

- Let $X =$ number of congenitally missing third molars in 20 year olds.
- $X$ can take the values: 0, 1, 2, 3, or 4.
- Each value has an associated probability.

<table>
<thead>
<tr>
<th>Probability Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
</tr>
<tr>
<td>$P(X = x)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Congenitally Missing Third Molars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability Distribution</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>
What is the probability of a 20-year old having two or more congenitally missing third molars?

<table>
<thead>
<tr>
<th>x</th>
<th>P(X = x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\[
P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4)
= 0.07 + 0.03 + 0.01
= 0.11
\]
Expected Value

• The **Expected Value** of a random variable $X$ is the value one would expect to get from taking the mean of an infinite number of observations of the random variable.

• Usually written $E(X)$

• $E(X)$ is also referred to as the “population mean” of $X$, and is often denoted by the symbol $\mu$ (“mu”).

\[ E(X) = \mu = \sum x \cdot P(X = x) \]
Example: congenitally missing third molars

<table>
<thead>
<tr>
<th>x</th>
<th>P(X = x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
</tr>
</tbody>
</table>

What is the expected number of congenitally missing third molars?

\[
E(X) = \sum x \cdot P(X = x)
\]

\[
= 0 \cdot 0.80 + 1 \cdot 0.09 + 2 \cdot 0.07 + 3 \cdot 0.03 + 4 \cdot 0.01
\]

\[
= 0.36
\]
Population Variance

- The **Population Variance** is the expected/average squared distance from the mean
- Denoted $\text{Var}(X)$ and $\sigma^2$
- Computation is similar to expected value

$$\text{Var}(X) = \sigma^2 = \sum (x - \mu)^2 \cdot P(X = x)$$
Population Standard Deviation

• The **Population Standard Deviation** is the square root of the population variance and is denoted by $\sigma$

\[
\sigma = \sqrt{Var(X)}
\]

• $\sigma$ is a measure of the spread of the population and can be loosely thought of as the average distance from the mean.

• In many cases 95% of the population will fall within $2 \sigma$ of the mean.
Example: calculation of population variance

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(X = x)</td>
<td>0.80</td>
<td>0.09</td>
<td>0.07</td>
<td>0.03</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\[E(X) = 0.036 \quad \text{(from previous slide)}\]

\[\text{Var}(X) = (0 - 0.36)^2 \cdot 0.80 + (1 - 0.36)^2 \cdot 0.09 + (2 - 0.36)^2 \cdot 0.07 + (3 - 0.36)^2 \cdot 0.03 + (4 - 0.36)^2 \cdot 0.01 = 0.116\]

\[\sigma = \sqrt{0.116} = 0.341\]
Binomial Distribution

• An especially useful discrete probability distribution.
• It is the count of the number of “successes” in a series of $n$ trials.
• The $n$ trials are assumed to be:
  • **Identical** - same probability of success, $p$, for each trial.
  • **Independent** - results of one trial does not influence a different trial.
Binomial Distribution

• The count of successes is an example of a case where we create a summary statistic by combining the results from a number of independent simple events.

• Working with such summary statistics and their distributions will be our focus in inferential statistics.
Example: Adverse drug reactions

• Rate of reactions for a certain drug is said to be 10%

• Dentist notes that of 3 patients she has prescribed the drug, 2 have experienced adverse reactions.

• Is this likely?

• If the probability is low, if may indicate that the reaction rate estimate of 10% may not be applicable to this population.
Example: Adverse drug reactions

- We can use the binomial distribution to calculate the probability of seeing 2 reactions out of 3 patients under the assumptions that:
  - the three patients are independent,
  - each has 10% chance of an adverse reaction.
Calculating the binomial probabilities

- Break up sample space into situations for which the probabilities can be calculated.

<table>
<thead>
<tr>
<th>Patient 1</th>
<th>X</th>
<th>X</th>
<th>X</th>
<th>X</th>
<th>-</th>
<th>-</th>
<th>-</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 2</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Patient 3</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Calculating the binomial probabilities

• Break up sample space into situations for which the probabilities can be calculated.

Possible outcomes (‘X’ = adverse reaction)

| Patient 1 | X | X | X | X | - | - | - | - | - | - |
| Patient 2 | X | X | - | - | - | X | X | - | - | - |
| Patient 3 | X | - | X | - | X | - | X | - | - | - |

Probability: .001

3 adverse events
P = .1 \times .1 \times .1
Calculating the binomial probabilities

- Break up sample space into situations for which the probabilities can be calculated.

### Possible outcomes (‘X’ = adverse reaction)

<table>
<thead>
<tr>
<th>Patient 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient 2</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient 3</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Probability</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>Probability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.009</td>
</tr>
</tbody>
</table>

3 adverse events
P = .1 × .1 × .1

2 adverse events
P = .1 × .1 × .9
Calculating the binomial probabilities

• Break up sample space into situations for which the probabilities can be calculated.

Possible outcomes (‘X’ = adverse reaction)

Patient 1  X  X  X  X  -  -  -  -  -
Patient 2  X  X  -  -  -  X  X  -  -
Patient 3  X  -  X  -  -  X  -  X  -

Probability  .001  .009  .009  .081  .009  .081  .081

3 adverse events  P = .1 × .1 × .1
2 adverse events  P = .1 × .1 × .9
1 adverse event  P = .1 × .9 × .9
Calculating the binomial probabilities

• Break up sample space into situations for which the probabilities can be calculated.

<table>
<thead>
<tr>
<th>Possible outcomes (‘X’ = adverse reaction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
</tr>
<tr>
<td>Patient 2</td>
</tr>
<tr>
<td>Patient 3</td>
</tr>
</tbody>
</table>

| Probability     | .001 | .009 | .009 | .081 | .009 | .081 | .081 | .729 |

3 adverse events  
P = .1 × .1 × .1

2 adverse events  
P = .1 × .1 × .9

1 adverse event   
P = .1 × .9 × .9

0 adverse events  
P = .9 × .9 × .9
Calculating the binomial probabilities

• Break up sample space into situations for which the probabilities can be calculated.

Possible outcomes (‘X’ = adverse reaction)

| Patient 1 | X | X | X | X | - | - | - | - |
| Patient 2 | X | X | - | - | X | X | - | - |
| Patient 3 | X | - | X | - | X | - | X | - |

Probability .001 .009 .009 .081 .009 .081 .081 .729
Calculating the binomial probabilities

• Now rearrange probabilities by number of adverse reactions

<table>
<thead>
<tr>
<th>Number of Patients</th>
<th>Probability</th>
<th>Cumulative Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1 \times 0.729 = 0.729$</td>
<td>0.729</td>
</tr>
<tr>
<td>1</td>
<td>$3 \times 0.081 = 0.243$</td>
<td>0.972</td>
</tr>
<tr>
<td>2</td>
<td>$3 \times 0.009 = 0.027$</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>$1 \times 0.001 = 0.001$</td>
<td>1.000</td>
</tr>
</tbody>
</table>

• This is a probability distribution

• Binomial distribution with $n = 3$ trials and probability of success (on each trial) of $p = 0.1$. 
Were 2 out of 3 adverse reactions likely?

<table>
<thead>
<tr>
<th>Number of Patients</th>
<th>Probability</th>
<th>Cumulative Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1 \times 0.729 = 0.729$</td>
<td>0.729</td>
</tr>
<tr>
<td>1</td>
<td>$3 \times 0.081 = 0.243$</td>
<td>0.972</td>
</tr>
<tr>
<td>2</td>
<td>$3 \times 0.009 = 0.027$</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>$1 \times 0.001 = 0.001$</td>
<td>1.000</td>
</tr>
</tbody>
</table>

- $P(\text{exactly 2 adverse reactions}) = 0.027$
- $P(\text{at least 2}) = 0.027 + 0.001 = 0.028$
- Thus, it was fairly unlikely event. Could lead one to believe that the 10% adverse event rate might not be correct
Graphical representation of probability distribution

binomial distribution n=3, p=0.1

- Number of successes
- Probability
Binomial distributions with different n’s and p’s

- $n = 6, \ p = 10\%$
- $n = 20, \ p = 10\%$
- $n = 80, \ p = 10\%$

- $n = 6, \ p = 50\%$
- $n = 20, \ p = 50\%$
- $n = 80, \ p = 50\%$
Computation formula for Binomial Probabilities

Suppose $X$ has binomial distribution with $n$ trials and success probability $p$, then

$$P(\text{exactly } k \text{ successes}) = \binom{n}{k} p^k (1 - p)^{n-k},$$

Where $\binom{n}{k}$, denoted “$n$ choose $k$”, is the number of ways one could choose $k$ items from a group of $n$ items.
“n choose k” examples and formula

\[ \binom{4}{2} = 6, \text{ as there are 6 ways to choose 2 items from a group of 4} \]

\[ \binom{n}{1} = n, \text{ there are } n \text{ ways to choose 1 item from a group of } n \]

\[ \binom{n}{0} = 1, \text{ only one way to choose no items} \]

In general, \( \binom{n}{k} = \frac{n!}{k!(n-k)!} \) where \( n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 \)

Note: \( 0! = 1 \)
Examples: Computation of binomial probabilities

If $X \sim \text{binomial}(n, p)$

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

Suppose a procedure has a 75% cure rate. What is the probability of 2 or fewer patients cured when the treatment is applied to 4 patients?

$X \sim \text{binomial}(n = 4, p = 0.75)$

$P(2 \text{ or fewer}) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$= 1 \cdot 0.75^0 \cdot 0.25^4 + 4 \cdot 0.75^1 \cdot 0.25^3 + 6 \cdot 0.75^2 \cdot 0.25^2$

$= 0.262$
Easier ways to find binomial probabilities

- Selected binomial distribution probabilities are presented in Table 1 on p. 169 in the coursepack (similar tables in Rosner appendix).

- Microsoft Excel can be used to calculate binomial probabilities using the function `BINOM.DIST`.

- The statistical software `R` can be used using the function `pbinom`.

- When $n$ is large one can also use the Normal distribution to compute approximate Binomial probabilities (see coursepack section 6.5).
For $X \sim \text{binomial}(n=20, p=0.05)$, $P(X = 2) = 0.1887$
Example: bronchitis

How likely are infants in at least 3 out of 20 households to develop chronic bronchitis if the probability of developing disease in any one household is 0.05?

The answer to this problem is $P(X \geq 3)$, where $X \sim \text{binomial}(n=20, \ p = .05)$.

Table 1 only gives us probabilities of $X$ being exactly equal to one number, \{P(X=3), P(X=4), etc.,\}, not P(X $\geq$ 3). So to compute P(X $\geq$ 3), we need to break into

$$P(X \geq 3) = P(X=3) + P(X=4) + \ldots + P(X=19) + P(X=20)$$

... which is a lot of work!
Example: bronchitis

Can use the complementary probability to reduce the calculations

\[ P(X \geq 3) = 1 - P(X \leq 2) \]
\[ = 1 - \{P(X=0) + P(X=1) + P(X=2)\} \]
\[ = 1 - (0.3585 + 0.3774 + 0.1887) \]
\[ = 1 - .9246 \]
\[ = 0.0754 \]
Computing binomial probabilities using Excel

• One can use the BINOM.DIST function in Excel to compute binomial probabilities.
• Can compute probabilities of the form $P(X = k)$
• Can also compute cumulative probabilities $P(X \leq k)$.
• The BINOM.DIST function has 4 arguments

$\text{BINOM.DIST}(\text{Number}_s, \text{Trials}, \text{Probability}_s, \text{Cumulative})$

• $\text{Number}_s$ is $k$
• $\text{Trials}$ is $n$
• $\text{Probability}_s$, is $p$
• $\text{Cumulative}$ tells whether you want a cumulative or an exact probability. TRUE gives $P(X \leq k)$, FALSE gives $P(X = k)$
Example:

• Compute the probability of 529 or fewer successes out of 1000 trials if the true probability of success on any trial is 50%.

• To compute $P(X \leq 529)$ where $X \sim \text{binomial}(n=1000, p=0.50)$ type into a cell:

  $$= \text{BINOM.DIST}(529, 1000, 0.50, \text{TRUE})$$

• Note: “FALSE” in the fourth argument would give $P(X = 529)$
Example:

=BINOM.DIST(529,1000,.5,TRUE)

Number_s 529 = 529
Trials 1000 = 1000
Probability_s .5 = 0.5
Cumulative TRUE = TRUE

Returns the individual term binomial distribution probability.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.968988402

Help on this function
Mean and Standard Deviation of Binomial

- For a binomial distribution $X$ with $n$ trials and probability of success $p$ on each trial, it can be worked out mathematically that:
  - Expected value, $E(X) = \mu = np$
  - $Var(X) = \sigma^2 = np(1-p)$
  - Standard Deviation $\sigma = \sqrt{np(1-p)}$
Example: drug effectiveness

• Suppose a certain drug has a 75% success rate. If you gave it to 100 patients then you would “expect” that the drug would work successfully on about

\[ np = 100 \cdot 0.75 = 75 \text{ cases.} \]

• Knowing the standard deviation can be useful because of the fact that usually (especially when \( n \) is large) 95% of the distribution should fall within \( 2\sigma \) of the mean.

\[ \sigma = \sqrt{np(1 - p)} = \sqrt{100 \cdot 0.75 \cdot 0.25} = 4.3 \]

• So while we’d expect about 75 people to be cured, we can be fairly certain (95% sure) that the number of successes would be within \( 2\times 4.3 \approx 9 \) of 75 (between 66 and 84 people).
One more example:

Suppose $X \sim \text{binomial}(n=14, p = .8)$, compute $P(X \leq 9)$.

- Because Table 1 does not have $p=.80$, we can solve the problem by converting the statement to refer to probability of failures rather than successes (prob of failure is .20).
- $P(9 \text{ or fewer successes}) = P(5 \text{ or more failures})$

Let $X^* = \text{number of failures}$, $X^* \sim \text{binomial}(n=14, p = .2)$

\[
P(X^* \geq 5) = 1 - P(X^* \leq 4)
= 1 - (P(X^*=0) + P(X^*=1) + ... + P(X^*=4))
= 1 - (.0440+.1539+.2501+.2501+.1720)
= 1 - .8701 = .129\]