Unit 3.

Discrete Distributions

<table>
<thead>
<tr>
<th>Topic</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Proportions and Rates in Epidemiological Research</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2.</td>
<td>Review - Bernoulli Distribution</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>4.</td>
<td>Poisson Distribution</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>5.</td>
<td>Hypergeometric Distribution</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>6.</td>
<td>Fisher’s Exact Test</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>7.</td>
<td>Discrete Distribution – Themes</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
</tbody>
</table>
1. Proportions and Rates in Epidemiological Research

The concepts of “proportion of” and “rate of” describe different aspects of disease occurrence. This distinction is important to epidemiological research.

A proportion is a relative frequency. It is dimensionless.

\[ \text{Proportion} = \frac{\# \text{ events that actually occurred}}{\# \text{ events that could have occurred}} \]

Valid range: 0 to 1

E.g. – Toss a coin 10 times. If we observe 2 “heads”:

Proportion “heads” = \( \frac{2}{10} = 20\% \)

- # Events that could have occurred = 10 tosses
- # Events occurred = 2 heads

Prevalence measures are examples of proportions.

A rate is a count of event occurrence per unit of time. As such, it is measured relative to an interval of time. It is not dimensionless.

\[ \text{Rate} = \frac{\# \text{ events that actually occurred}}{\# \text{ time periods experienced}} \]

Valid range: 0 to \( \infty \)

E.g., - 100 persons are known to have smoked, collectively, for 1,000 pack years. If we observe 3 occurrences of lung cancer:

Rate lung cancer = \( \frac{3}{1000} \) pack years

- # Time periods experienced = 1000 pack years
- # Events occurred = 3

Incidence densities are examples of rates.
Some Commonly Used “Proportions” and “Rates”

Proportions describe either existing disease or new disease within a time frame. Rates describe the “force” or “flow” of occurrence of new disease with time.

Prevalence = \frac{\text{# persons with disease at a point in time}}{\text{# persons in the population at a point in time}}

<table>
<thead>
<tr>
<th>Prevalence</th>
<th>Type of Measure: Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denominator = # events that could have occurred</td>
<td># persons in the population at a point in time</td>
</tr>
<tr>
<td>Numerator = # events that actually occurred</td>
<td># persons having the disease at a point in time</td>
</tr>
<tr>
<td>Valid range:</td>
<td>0 to 1</td>
</tr>
<tr>
<td>Interpretation</td>
<td>The proportion of the population with disease at a point in time</td>
</tr>
<tr>
<td>Other names</td>
<td>Point prevalence</td>
</tr>
</tbody>
</table>

Example:

In 1988, the New York State Breast and Cervical Cancer Screening Program registry included 16,529 women with a baseline mammogram.

Upon review of their medical records, 528 were found to have a history of breast cancer.

The prevalence of breast cancer in this 1988 selected cohort was

\[ P = \frac{528}{16,529} = 0.0319 = 3.19\% \]

These 528 women were excluded from the analyses to determine the factors associated with repeat screening mammogram.
Cumulative incidence = \( \frac{\text{# disease onsets during interval of time}}{\text{# persons at risk in population at start of interval}} \)

<table>
<thead>
<tr>
<th>Cumulative Incidence</th>
<th>Type of Measure: Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denominator = # events that could have happened</td>
<td># persons in the population at the start of the time interval who could possibly develop disease. Thus, all are disease free at the start of the interval.</td>
</tr>
<tr>
<td>Numerator = # events that actually occurred</td>
<td># persons developing the disease during the time interval.</td>
</tr>
<tr>
<td>Valid range:</td>
<td>0 to 1</td>
</tr>
<tr>
<td>Interpretation</td>
<td>The proportion of healthy persons who go on to develop disease over a specified time period.</td>
</tr>
<tr>
<td>Note !!</td>
<td>We assume that every person in the population at the start was followed for the entire interval of time.</td>
</tr>
</tbody>
</table>

Example:
In this example, the event of interest is the completion of a repeat screening mammogram.

There were 9,485 women in the New York State Breast and Cervical Cancer Screening Program registry as of 1988 with a negative mammogram, who did not have a history of breast cancer, and provided complete data.

2,604 obtained a repeat screening mammogram during the 6 year period 1988-1993.

The 6-year cumulative incidence of repeat screening mammogram is therefore

\[
CI_{6\text{year}} = \frac{2,604}{9,485} = 0.2745 = 27%
\]

Further analysis is focused on the identification of its correlates.
**Incidence Density** = \( \frac{\text{# disease onsets during interval of time}}{\text{Sum of individual lengths of time actually at risk}} \)

<table>
<thead>
<tr>
<th>Incidence Density</th>
<th>Type of Measure: Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Denominator</strong> = # time periods experienced event free.</td>
<td>Sum of individual lengths of time during which there was opportunity for event occurrence during a specified interval of time.</td>
</tr>
<tr>
<td>This sum is called “person time” and is ( \sum_{i=1}^{N} (\text{time for } i^{th} \text{ person free of event}) ).</td>
<td>It is also called “person years” or “risk time”.</td>
</tr>
<tr>
<td><strong>Numerator</strong> = # new event occurrences</td>
<td># disease onsets during a specified interval of time.</td>
</tr>
<tr>
<td>Valid range:</td>
<td>0 to ( \infty )</td>
</tr>
<tr>
<td>Interpretation</td>
<td>The force of disease occurrence per unit of time</td>
</tr>
<tr>
<td>Other names</td>
<td>Incidence rate</td>
</tr>
<tr>
<td>Note!!</td>
<td>We must assume that the “risk” of event remains constant over time. When it doesn’t, stratified approaches are required.</td>
</tr>
</tbody>
</table>

*Be careful in the reporting of an incidence density.*

*Don’t forget the scale of measurement of time.*

\[
I = \text{7 per “person year”}
\]

\[
= \frac{7}{52} = 0.13 \text{ per “person week”}
\]

\[
= \frac{7}{365.25} = 0.019 \text{ per “person day”}
\]
What to Use?

A prevalence estimate is useful when interest is in

- who has disease now versus who does not (one time camera picture)
- planning services; e.g. - delivery of health care

The concept of prevalence is not meaningfully applicable to etiologic studies.

- Susceptibility and duration of disease contribute to prevalence
- Thus, prevalence = function(susceptibility, incidence, survival)

Etiologic studies of disease occurrence often use the cumulative incidence measure of frequency.

- Recall that we must assume complete follow-up of entire study cohort.

The cumulative incidence measure of disease frequency is not helpful to us if persons migrate in and out of the study population.

- Individuals no longer have the same opportunity for event recognition.

Etiologic studies of disease occurrence in dynamic populations will then use the incidence density measure of frequency.

- Be careful here, too! Does risk of event change with time? With age?

- If so, calculate person time separately in each of several “blocks” of time. This is a stratified analysis approach.
2. Review - Bernoulli Distribution

In order to analyze variations in proportions and rates, we’ll need a few probability models. There are 4 and they are interrelated.

Two of them were introduced in PubHlth 540, Introductory Biostatistics:

- Bernoulli; and
- Binomial.

Two additional distributions that are often used in analyses of discrete data are the following:

- Poisson; and
- hypergeometric.

**Recall** - A Bernoulli Trial is the Simplest Binomial Random Variable.

A simple example is the coin toss.

“50-50 chance of heads” can be re-cast as a random variable. Let

\[
Z = \text{random variable representing outcome of one toss, with}
\]

\[
Z = 1 \text{ if “heads”}
\]

\[
0 \text{ if “tails”}
\]

\[
\pi = \text{Probability [ coin lands “heads” ]}. \text{ Thus,}
\]

\[
\pi = \text{Probability [ } Z = 1 \text{ ]}
\]
We have what we need to define a probability distribution.

<table>
<thead>
<tr>
<th>Enumeration of all possible outcomes</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>- outcomes are mutually exclusive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- outcomes exhaust all possibilities</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Associated probabilities of each</th>
<th>Outcome</th>
<th>Pr[outcome]</th>
</tr>
</thead>
<tbody>
<tr>
<td>- each probability is between 0 and 1</td>
<td>1</td>
<td>(\Pr[Z=1] = \pi)</td>
</tr>
<tr>
<td>- sum of probabilities totals 1</td>
<td>0</td>
<td>(\Pr[Z=0] = (1-\pi))</td>
</tr>
<tr>
<td></td>
<td>Total = 1</td>
<td></td>
</tr>
</tbody>
</table>

**In epidemiology, the Bernoulli might be a model for the description of ONE individual (N=1):**

This person is in one of two states. He or she is either in a state of:

1) “event” with probability \(\pi\); or
2) “non event” with probability \((1-\pi)\)

The description of the likelihood of being either in the “event” state or the “non-event” state is given by the **Bernoulli distribution**
**Bernoulli Distribution**

Suppose Z can take on only two values, 1 or 0, and suppose:

\[
\text{Probability } [ Z = 1 ] = \pi \\
\text{Probability } [ Z = 0 ] = (1-\pi)
\]

This gives us the following expression for the likelihood of Z=z.

\[
\text{Probability } [ Z = z ] = \pi^z (1-\pi)^{1-z} \text{ for } z = 0 \text{ or } 1.
\]

Expected value of Z is \(E[Z] = \pi\)

Variance of Z is \(\text{Var}[Z] = \pi (1-\pi)\)

**Example**: Z is the result of tossing a coin once. If it lands “heads” with probability = .5, then \(\pi = .5\).

Later we’ll see that individual Bernoulli distributions are the basis of describing patterns of disease occurrence in a **logistic regression** analysis.
**Mean (μ) and Variance (σ²) of a Bernoulli Distribution**

Mean of Z = μ = π

The mean of Z is represented as E[Z].  

\[ E[Z] = \pi \]

because the following is true:

\[
E[Z] = \sum_{\text{All possible } z} [z] \text{Probability}[Z=z]
= [0] \Pr[Z=0] + [1] \Pr[Z=1]
= [0] (1-\pi) + [1] (\pi)
= \pi
\]

Variance of Z = σ² = (π)(1-π)

The variance of Z is Var[Z] = E[ (Z – (E[Z])² ].

\[ \text{Var}[Z] = \pi(1-\pi) \]

because the following is true:

\[
\text{Var}[Z] = E [(Z-\pi)^2] = \sum_{\text{All possible } z} (z-\pi)^2 \text{Probability}[Z=z]
= [(0-\pi)^2]\Pr[Z=0]+[(1-\pi)^2]\Pr[Z=1]
= [\pi^2] (1-\pi) + [(1-\pi)^2] (\pi)
= \pi (1-\pi) [ \pi + (1-\pi) ]
= \pi (1-\pi)
\]
3. Review - Binomial Distribution

There are at least two ways to think about the Binomial distribution:

- 1 - A Binomial outcome is actually the sum of N Bernoulli trials
- 2 - Generically

1 - The Binomial as the Sum of Several Bernoulli Trials.

The example of tossing one coin one time is an example of 1 Bernoulli trial. Suppose we call this random variable Z:

- \( Z \) is distributed Bernoulli (\( \pi \)) with \( Z=1 \) when the event occurs, 0 when it does not.

If we toss the same coin several times, say N times, we have N Bernoulli trials:

- \( Z_1, Z_2, \ldots, Z_N \) are each distributed Bernoulli (\( \pi \)) and they’re independent.

Consider what happens if we just add up the \( Z \)'s. We’re actually adding up 1’s and 0’s. The total of \( Z_1, Z_2, \ldots, Z_N \) is thus the total number of 1’s. It is also the net number of events of success in N trials. Let’s call this number of events of success a new random variable \( X \).

\[ \sum_{i=1}^{N} Z_i = X = \text{# events in N trials} \]

This new random variable \( X \) is distributed Binomial. It represents the net number of successes in a set of independent Bernoulli trials. A simple example is the outcome of several coin tosses (eg – “how many heads did I get?”). The word choice “net” is deliberate here, to remind ourselves that we’re not interested in keeping track of the particular trials that yielded events of success, only the net number of trials that yielded event of success.

E.g.

- What is the probability that 2 of 6 graduate students are female?
- What is the probability that of 100 infected persons, 4 will die within a year?
Steps in calculating the probability of $\sum_{i=1}^{N} Z_i = X = x$

**Step 1** – Pick just one arrangement of x events in N trials and calculate its probability

The easiest is the ordered sequence consisting of (x) events followed by (N-x) non events.

\[
\begin{array}{cccccccc}
1 & 1 & 1 & \ldots & 1 & 0 & 0 & 0 & \ldots & 0 \\
\end{array}
\]

\[\text{x events} \quad \text{N-x non events}\]

\[
\Pr \text{ [ ordered sequence ] } = \Pr \left[ (Z_1 = 1), (Z_2 = 1) \ldots (Z_x = 1), (Z_{x+1} = 0), (Z_{x+2} = 0) \ldots (Z_N = 0) \right]
\]

\[= \pi \pi \pi \ldots \pi (1-\pi) (1-\pi) (1-\pi) \ldots (1-\pi)\]

\[= \pi^x (1-\pi)^{N-x}\]

**Note!** Can you see that the result is the same product $\pi^x (1-\pi)^{N-x}$ regardless of where in the sequence the x events occurred? How handy!

**Step 2** – Determine the number of "qualifying" ordered sequences that satisfy the requirement of having exactly x events and (N-x) non events.

Number of "qualifying" ordered sequences = \[\binom{N}{x} = \frac{N!}{x! (N-x)!}\]

**Step 3** – The probability of getting X=x events, without regard to sequencing, is thus the sum of the probabilities of each “qualifying” ordered sequence, all of which have the same probability that was obtained in step 1.

Probability [N trials yields x events] = (# qualifying sequence) (Pr[one sequence])

\[\Pr[X = x] = \Pr \left[ \sum_{i=1}^{N} Z_i = x \right] = \binom{N}{x} \pi^x (1-\pi)^{N-x}\]
2 - The Binomial in General Form

The Binomial Distribution

Among N trials, where

- The N trials are independent
- Each trial has two possible outcomes: 1 or 0
- \( \Pr[\text{outcome}=1] = \pi \) for each trial; and therefore \( \Pr[\text{outcome}=0] = (1-\pi) \)
- \( X = \# \) events of outcome

\[
\text{Probability} [ X = x ] = \binom{N}{x} \pi^x (1-\pi)^{N-x} \quad \text{for } x=0, \ldots, N
\]

Expected value is \( E[\sum_{i=1}^{N} Z_i = x] = N \pi \)

Variance is \( \text{Var}[\sum_{i=1}^{N} Z_i = x] = N \pi (1-\pi) \)

\[
\binom{N}{x} = \# \text{ ways to choose } X \text{ from } N = \frac{N!}{x! (N-x)!}
\]

where \( N! = N(N-1)(N-2)(N-3) \cdots (4)(3)(2)(1) \) and is called the factorial.
Your Turn

A roulette wheel lands on each of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 with probability = .10. Write down the expression for the calculation of the following.

#1. The probability of “5 or 6” exactly 3 times in 20 spins.

#2. The probability of “digit greater than 6” at most 3 times in 20 spins.
#1. **Solution:**

“Event” is outcome of either “5” or “6”

\[
\Pr[\text{event}] = \pi = 0.20
\]

\[N = 20\]

\[X \text{ is distributed } \text{Binomial}(N=20, \pi=0.20)\]

\[\Pr[X=3] = \binom{20}{3} \cdot [0.20]^3 \cdot [1-0.20]^{20-3}\]

\[= \binom{20}{3} \cdot [0.20]^3 \cdot [0.80]^{17}\]

\[= 0.2054\]  

#2. **Solution:**

“Event” is outcome of either “7” or “8” or “9”

\[\Pr[\text{event}] = \pi = 0.30\]

\[N = 20\]

\[X \text{ is distributed } \text{Binomial}(N=20, \pi=0.30)\]

*Translation:* “At most 3 times” is the same as saying “3 times or 2 times or 1 time or 0 times” which is the same as saying “less than or equal to 3 times”

\[\Pr[X \leq 3] = \Pr[X=0] + \Pr[X=1] + \Pr[X=2] + \Pr[X=3]\]

\[= \sum_{x=0}^{3} \binom{20}{x} \cdot [0.30]^x \cdot [0.70]^{20-x}\]

\[= \binom{20}{0} \cdot [0.30]^0 \cdot [0.70]^{20} + \binom{20}{1} \cdot [0.30]^1 \cdot [0.70]^{19} + \binom{20}{2} \cdot [0.30]^2 \cdot [0.70]^{18} + \binom{20}{3} \cdot [0.30]^3 \cdot [0.70]^{17}\]

\[= 0.10709\]
Review of the Normal Approximation for the Calculation of Binomial Probabilities

Calculations of exact binomial probabilities become quite tedious as the number of trials and number of events gets large. Fortunately, the central limit theorem allows us to replace these exact calculations with very good approximate calculations. The approximate calculations are actually normal probability calculations.

The following is an example where the calculation of the required binomial probability would be too much to do!

Example
Calculate the chances of between 5 and 28 events (inclusive) in 180 trials with probability of event = .041.

Idea of Solution
Translate the required exact calculation into a very good approximate one using the z-score.

X distributed Binomial (N=180, π=.041) says that

\[ \mu_{\text{BINOMIAL}} = n\pi = (180)(.041) = 7.38 \]
\[ \sigma^2_{\text{BINOMIAL}} = n\pi(1-\pi) = (180)(.041)(.959) = 7.08 \]
\[ \sigma_{\text{BINOMIAL}} = \sqrt{\sigma^2_{\text{BINOMIAL}}} = 2.66 \]

The approximate calculation using the z-score uses \( \mu = \mu_{\text{BINOMIAL}} \) and \( \sigma = \sigma_{\text{BINOMIAL}} \)

\[ \Pr[5 \leq X \leq 28] \approx \Pr\left[ \frac{5-\mu}{\sigma} \leq \text{Normal}(0,1) \leq \frac{28-\mu}{\sigma} \right] \]

\[ = \Pr\left[ \frac{5 - 7.38}{2.66} \leq \text{Normal}(0,1) \leq \frac{28 - 7.38}{2.66} \right] \]

\[ = \Pr[ -0.895 \leq \text{Normal}(0,1) \leq 7.752 ] \]

\[ \approx \Pr[ -0.895 \leq \text{Normal}(0,1) ] , \text{because 7.752 is in the extreme right tail.} \]

\[ = \Pr[ \text{Normal (0,1)} \leq +0.895 ] , \text{because of symmetry of the tails of the normal} \]

\[ = .8146 \]
Review of the Normal Approximation for the Calculation of Binomial Probabilities – continued.

More generally, we can use a z-score and the normal distribution for the following reasons.

- Binomial probabilities are likelihood calculations for a discrete random variable. Normal distribution probabilities are likelihood calculations for a continuous random variable.

- When substituting for the exact probabilities, we use the Normal distribution that has mean and variance parameter values equal to those of our Binomial distribution.

\[
\mu_{\text{normal}} = \mu_{\text{binomial}} = n\pi \\
\sigma^2_{\text{normal}} = \sigma^2_{\text{binomial}} = n\pi (1-\pi)
\]

<table>
<thead>
<tr>
<th>Desired Binomial Probability Calculation</th>
<th>Normal Approximation w Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr [ X=k ]</td>
<td>Pr [ (k-1/2) \leq X \leq (k+1/2) ]</td>
</tr>
<tr>
<td>Pr [ X &gt; k ]</td>
<td>Pr [ X &gt; (k-1/2) ]</td>
</tr>
<tr>
<td>Pr [ X &lt; k ]</td>
<td>Pr [ X &lt; (k+1/2) ]</td>
</tr>
</tbody>
</table>
4. Poisson Distribution

So far we have considered

- The description of 1 event occurrence for 1 person (Bernoulli)
- The description of X event occurrences in sample of N persons (Binomial)

Instead of thinking in terms of persons, think instead about PERSON TIME:

A familiar example of the idea of person time is “pack years smoking”

E.g. – How shall we describe a small number of cancer deaths relative to a large accumulation of person time, such as 3 cancer deaths in 1000 pack years of smoking?

**Setting.**

- It is no longer the analysis of events among N persons as a proportion.
- Instead, it is an analysis of events over person time as an incidence rate.

The Poisson Distribution can be appreciated as an extension of the Binomial Distribution.

- The concept of N persons → A large accumulation of person time
- The likelihood of an event experienced by 1 person →
  the likelihood of an event in 1 unit of person time.
  This will be quite small!
The extension.

- We begin by constructing a binomial likelihood situation. Let

  \[ T = \text{total accumulation of person time} \quad (\text{e.g.} - 1000 \text{ pack years}) \]
  \[ n = \text{number of sub-intervals of } T \quad (\text{e.g.} - 1000) \]
  \[ T/n = \text{length of 1 sub-interval of } T \quad (\text{e.g.}- 1 \text{ pack year}) \]
  \[ \lambda = \text{event rate per unit length of person time} \]

- What is our Binomial distribution probability parameter \( \pi \)?

  \[ \pi = \lambda (T/n) \text{ because it is (rate)(length of 1 sub-interval)} \]

- What is our Binomial distribution number of trials?

  \[ n = \text{number of sub-intervals of } T \]

We’ll need 3 assumptions

1) The rate of events in each sub-interval is less than 1.

\[
\text{Rate per sub-interval} = \Pr[1 \text{ event per sub-interval}]
\]

\[
0 < \pi = (\lambda)(T/n) < 1
\]

2) The chances of 2 or more events in a sub-interval is zero.

3) The subintervals are mutually independent.

Now we can describe event occurrence over the entire interval of length \( T \) with the Binomial.

Let \( X \) be the count of number of events.

Probability \( [X = x] \) = \( \binom{n}{x} \pi^x (1-\pi)^{n-x} \) for \( x=0, \ldots, n \)

\[
= \binom{n}{x} \left( \frac{\lambda T}{n} \right)^x \left[ 1 - \frac{\lambda T}{n} \right]^{n-x} \because \pi = (\lambda)(T/n).
\]
Some algebra (if you care to follow along) will get us to the Poisson distribution probability formula.

The algebra involves two things:

- Letting $n \to \infty$ in the binomial distribution probability; and

- Recognizing that the expected number of events over the entire interval of length $T$ is $\lambda T$ because $\lambda$ is the “per unit subinterval” rate and $T$ is the number of units. (analogy: for rate of heads $= .50$, number of coin tosses $= 20$, the expected number of head is $[.50][20] = 10$). This allows us, eventually, to make the substitution of $\lambda T = \mu$.

$$Pr[X=x] = \binom{n}{x} \left( \frac{\lambda T}{n} \right)^x \left[ 1 - \frac{\lambda T}{n} \right]^{n-x}$$

$$= \binom{n}{x} \left[ \frac{\lambda T}{n} \right]^x \left[ 1 - \frac{\lambda T}{n} \right]^n \left[ 1 - \frac{\lambda T}{n} \right]^{-x}$$

Work with each term on the right hand side one at a time:

**- 1st term -**

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n(n-1)(n-2)\ldots(n-x+1)(n-x)!}{x!(n-x)!} = \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!}$$

As $n \to \infty$, the product of terms in the numerator $\to (n)(n)(n) \ldots (n) = n^x$.

Thus, as $n \to \infty$

$$\binom{n}{x} \to \frac{n^x}{x!}$$
- 2\textsuperscript{nd} term -
\[
\left[ \frac{\lambda T}{n} \right]^x = [\lambda T]^x \left( \frac{1}{n} \right)^x = [\mu]^x \left( \frac{1}{n} \right)^x
\]

Thus,
\[
\left[ \frac{\lambda T}{n} \right]^x = [\mu]^x \left( \frac{1}{n} \right)^x
\]

- 3\textsuperscript{rd} term -
\[
\left[ 1 - \frac{\lambda T}{n} \right]^n = \left[ 1 - \frac{\mu}{n} \right]^n
\]

What happens next is a bit of calculus. As \( n \to \infty \)
\[
\left( \frac{1 - \frac{\mu}{n}}{n} \right)^n \to e^{-\mu}
\]

where \( e \) = constant \( \approx 2.718 \)

Thus, as \( n \to \infty \)
\[
\left[ 1 - \frac{\lambda T}{n} \right]^n = \left[ 1 - \frac{\mu}{n} \right]^n \to e^{-\mu}
\]

- 4\textsuperscript{th} term -

Finally, as \( n \to \infty \) the quotient \( \left( \frac{\lambda T}{n} \right) \) is increasingly like \( \left( \frac{0}{n} \right) \) so that
\[
\left[ 1 - \frac{\lambda T}{n} \right]^x \to \left[ 1 - \frac{0}{n} \right]^x \to [1]^x = 1
\]

Thus, as \( n \to \infty \)
\[
\left[ 1 - \frac{\lambda T}{n} \right]^x \to 1
\]
Now put together the product of the 4 terms and what happens as \( n \to \infty \)

\[
\Pr[X=x] = \binom{n}{x} \left(\frac{\lambda T}{n}\right)^x \left(1 - \frac{\lambda T}{n}\right)^n \left(1 - \frac{\lambda T}{n}\right)^x
\]

as \( n \to \infty \)

\[
\begin{align*}
\frac{n^x}{x!} &\quad \{ \mu \}^x \left[ \frac{1}{n} \right]^x \quad \{ e^{-\mu} \} \quad \{ 1 \} \\
&= \frac{\mu^x e^{-\mu}}{x!} = \frac{\mu^x \exp(-\mu)}{x!}
\end{align*}
\]

### Poisson Distribution

If \( X \) is distributed Poisson,

\[
\text{Probability } [ X = x ] = \frac{\mu^x \exp(-\mu)}{x!} \quad \text{for } x = 0, 1, \ldots, \infty
\]

Expected value of \( X \) is \( \text{E}[X] = \mu \)

Variance of \( X \) is \( \text{Var}[X] = \mu \)

The poisson distribution is an appropriate model for describing the frequency of occurrence of a rare event in a very large number of trials.
Example:

This example illustrates the correspondence between the Binomial and Poisson likelihoods. If lung cancer occurs at a rate of 2 per 1000 pack years, calculate the probability of exactly 3 cases of lung cancer in 3000 pack years using (a) a Binomial model, and (b) a Poisson model.

<table>
<thead>
<tr>
<th></th>
<th>Binomial</th>
<th>Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = # trials</td>
<td>$\pi = \Pr[\text{event}]$</td>
<td>What happens as $n \to \infty$ and $\pi \to 0$?</td>
</tr>
<tr>
<td>Expected # events</td>
<td>$n \pi$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>$\Pr[\text{X=x events}]$</td>
<td>$\binom{n}{x} \pi^x (1-\pi)^{n-x}$</td>
<td>$\frac{\mu^x \exp(-\mu)}{x!}$</td>
</tr>
</tbody>
</table>

Where $\exp = \text{numerical constant} = e = 2.718$

**Solution using the Binomial:**

# trials = $n = 3000$
$\pi = .002$

$\Pr[3 \text{ cases cancer}] = \binom{3000}{3} [.002]^3 [.998]^{2997} = .0891$

**Solution using the Poisson:**

Length of interval, $T = 3000$
Rate per unit length, $\lambda = .002$ per 1 pack year
$\mu = \lambda T = (.002/\text{pack year})(3000 \text{ pack years}) = 6$

$\Pr[3 \text{ cases cancer}] = \frac{6^3 \exp(-6)}{3!} = .0892$
An example of the Poisson sampling design is a cross-sectional study

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
<td>b+d</td>
</tr>
<tr>
<td></td>
<td>a+b+c+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

- The counts a, b, c, and d are each free to vary and follow their own distribution.
- a, b, c, and d are 4 independent Poisson random variables.
- Let’s denote the means of these $\mu_{11}$, $\mu_{12}$, $\mu_{21}$, and $\mu_{22}$.

Likelihood[2x2 table]

$$L[a, b, c, d] = \frac{\mu_{11}^a \exp(-\mu_{11})}{a!} \frac{\mu_{12}^b \exp(-\mu_{12})}{b!} \frac{\mu_{21}^c \exp(-\mu_{21})}{c!} \frac{\mu_{22}^d \exp(-\mu_{22})}{d!}$$

An association between exposure and disease means the following.

1) $Pr[\text{disease given exposure}] \neq Pr[\text{disease given no exposure}]
   \frac{\mu_{11}}{\mu_{11} + \mu_{12}} \neq \frac{\mu_{21}}{\mu_{21} + \mu_{22}}$

2) $Pr[\text{exposure given disease}] \neq Pr[\text{exposure given no disease}]
   \frac{\mu_{11}}{\mu_{11} + \mu_{12}} \neq \frac{\mu_{12}}{\mu_{12} + \mu_{22}}$
An example of the Binomial sampling design is a cohort study

| Exposed | Disease | Healthy | | | |
|---------|---------|---------|---|---|
| a       | b       | a+b fixed by design |
| c       | d       | c+d fixed by design |
| a+c     | b+d     | a+b+c+d |

- The counts “a” and “c” are each free to vary and follow their own distribution.
- The counts b and d do not vary because of the fixed row totals
- a and c are 2 independent Binomial random variables.
- Let’s denote the means of these $\pi_1$ and $\pi_2$.

Likelihood[2x2 table]

\[
L[a,c \text{ given } (a+b) \text{ and } (c+d) \text{ are fixed}]
\]

\[
= \left[ \binom{a+b}{a} \pi_1^a (1-\pi_1)^b \right] \left[ \binom{c+d}{c} \pi_2^c (1-\pi_2)^d \right]
\]

An association between exposure and disease means the following.
Pr[disease among exposed] $\neq$ Pr [disease among non-exposed]

\[
\pi_1 \neq \pi_2
\]
An example of the Binomial sampling design is a \textit{case-control} study

<table>
<thead>
<tr>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a \hspace{1cm} B</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c \hspace{1cm} D</td>
</tr>
</tbody>
</table>

\begin{itemize}
  \item a = \# persons with disease whose recall reveals exposure
  \item c = \# persons with disease whose recall reveals no exposure
  \item b = \# healthy persons whose recall reveals exposure
  \item d = \# healthy persons whose recall reveals no exposure
\end{itemize}

\begin{itemize}
  \item The counts a and b are each free to vary and follow their own distribution.
  \item The counts c and d do not vary because of the fixed column totals
  \item a and b are 2 independent Binomial random variables.
  \item Let’s denote the means of these \(\theta_1\) and \(\theta_2\).
\end{itemize}

Likelihood[2x2 table]

\[
L[a,b \text{ give } (a+c) \text{ and } (b+d) \text{ are fixed}] = \left[\frac{a+c}{a}\right]^{\theta_1(1-\theta_1)^c} \left[\frac{b+d}{b}\right]^{\theta_2(1-\theta_2)^d} \]

An association between exposure and disease means the following,
\[
\Pr[\text{exposure among disease}] \neq \Pr[\text{exposure among healthy}]
\]
\[
\theta_1 \neq \theta_2
\]
5. Hypergeometric Distribution

Sometimes, a 2x2 table analysis of an exposure-disease relationship has as its only focus the investigation of association. This will become more clear in units 4 and 5. In this setting, it turns out that the most appropriate probability model is the **hypergeometric distribution model**.

The Hypergeometric Distribution is introduced in two ways:

- 1 - The Hypergeometric Distribution in Games of Poker
- 2 - The Hypergeometric Distribution in Analyses of 2x2 Tables

**1 - The Hypergeometric Distribution in Games of Poker.**

**Example - In a five card hand, what is the probability of getting 2 queens?**

- 2 queens is, in reality, 2 queens and 3 NON queens.
- We know the total number of possible hands is \( \binom{52}{5} \)
- Because there are 4 queens in a full deck, the # selections of 2 queens from 4 is \( \binom{4}{2} \)
- Similarly, there are \( 52 - 4 = 48 \) “non” queens in a full deck. Thus, the # selections of 3 “non” queens from 48 is \( \binom{48}{3} \)
- Putting these together, the number of hands that have 2 queens and 3 “non queens is the product \( \binom{4}{2} \binom{48}{3} \)

\[
\text{Pr[2 queens in five card hand]} = \frac{(\text{Number of five card hands that have 2 queens})}{(\text{Total number of five card hands})} = \frac{\binom{4}{2} \binom{48}{3}}{\binom{52}{5}}
\]
2 - The Hypergeometric Distribution in Analyses of a 2x2 Table.

Example – A biotech company has N=259 pregnant women in its employ. 23 of them work with video display terminals. Of the 259 pregnancies, 4 ended in spontaneous abortion. Assuming a central hypergeometric model, what is the probability that 2 of the 4 spontaneous abortions were among the 23 women who worked with video display terminals?

<table>
<thead>
<tr>
<th>Video Display Terminal</th>
<th>Spontaneous Abortion</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>255</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>236</td>
</tr>
</tbody>
</table>

- In this example, the number of women who worked with video display terminals (23) is fixed. That is, the row totals are fixed.

- Similarly, the total number of spontaneous abortions (4) is also fixed. That is, the column totals are fixed.

- And the total number of pregnancies (259) is fixed.

- Thus, only one cell count can vary. Having chosen one, all others are known by subtraction.

- In this example, we want to know if spontaneous abortion has occurred disproportionately often among the video display terminal workers? Is it reasonable that 2 of the 4 cases of spontaneous abortion occurred among the relatively small number (23) of women who worked with video display terminals?

- The “chance” model is the central hypergeometric model just introduced. Using the approach just learned,
Example – At this biotech company, what is the probability that 2 of the 4 abortions occurred among the 23 who worked with video display terminals?

- 2 abortions among VDT workers is in reality, 2 abortions among VDT workers and 2 abortions among NON-VDT workers.

- We know the total number of possible choices of 4 abortions from 259 pregnancies is \( \binom{259}{4} \)

- As there are 23 VDT workers, the # selections of 2 from this group is \( \binom{23}{2} \)

- Similarly, as there are \( 259 - 23 = 236 \) non-VDT workers, the number of selections of 2 from this group is \( \binom{236}{2} \)

- Thus, the probability of 2 of the 4 abortions occurring among VDT workers by chance is the hypergeometric probability

\[
\frac{\binom{23}{2} \binom{236}{2}}{\binom{259}{4}}
\]
3 - A Hypergeometric Probability Model for a 2x2 Table of Exposure-Disease Counts

Now we consider the general setting of a 2x2 table analysis of exposure-disease count data.

<table>
<thead>
<tr>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
</tr>
<tr>
<td>(a+c) fixed</td>
<td>(b+d) fixed</td>
</tr>
<tr>
<td>n = a + b + c + d</td>
<td></td>
</tr>
</tbody>
</table>

Here, the **only** interesting cell count is

“a” = # persons with both exposure and disease

**Example –**

In this 2x2 table with fixed total exposed, and fixed total number of events, what is the probability of getting “a” events of (exposed and event)?

- “a” events of case among “(a+b)” exposed is in reality, “a” events among “(a+b)” exposed AND “c” events among “(c+d)” NOT exposed.

- We know the total number of possible choices of “(a+c)” cases from “(a+b+c+d)” total is \( \binom{a+b+c+d}{a+c} \)

- As there are “(a+b)” exposed, the # selections of “a” from this group is \( \binom{a+b}{a} \)

- Similarly, as there are “(c+d)” NOT exposed, the number of selections of “c” from this group is \( \binom{c+d}{c} \)

- Thus, the probability of “a” events of CASE occurring among “(a+c)” EXPOSED by chance is the **hypergeometric** probability

\[
\frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}
\]
Note – This is actually the central hypergeometric. More on this later….

Nice Result –

The central hypergeometric probability calculation for the 2x2 table is the same regardless of arrangement of rows and columns.

<table>
<thead>
<tr>
<th>Cohort Design</th>
<th>Case-Control Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr[&quot;a&quot; cases among &quot;(a+b)&quot; exposed</td>
<td>fixed marginals]</td>
</tr>
</tbody>
</table>
| \[
\frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}
\] | \[
\frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{a+b+c+d}{a+b}}
\] |
6. Fisher’s Exact Test of Association in a 2x2 Table

Now we can put this all together. We have a chance model for the probability of obtaining whatever count “a” we get in the upper left cell of the 2x2 table. It is the hypergeometric probability distribution model.

This “chance” model (the central hypergeometric distribution) is the null hypothesis probability model that is used in the Fisher’s exact test of no association in a 2x2 table.

The odds ratio OR is a single parameter which describes the exposure disease association.

Consider again the VDT exposure and occurrence of spontaneous abortion data:

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>Not exposed</td>
<td>2</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>255</td>
</tr>
</tbody>
</table>

We’re not interested in the row totals. Nor are we interested in the column totals. Thus, neither the Poisson nor the Bionomial likelihoods are appropriate models for our particular question.

Rather, our interest is in the number of persons who have both traits – (exposure=yes) and (disease=yes).

Is the count of 2 with exposure and disease significantly larger than what might have been expected if there were NO association between exposure and disease?

We’re interested in this likelihood because it is the association, the odds ratio OR, that is of interest, not the row totals, nor the column totals. Thus, we will take advantage of the central hypergeometric model.
**What is the null hypothesis likelihood of the 2x2 table if the row and column totals are fixed?**

Recall the layout of our 2x2 table. With row and column totals fixed, only one cell count can vary. Let this be the count “a”.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
</tr>
</tbody>
</table>

### Null Hypothesis Conditional Probability of 2x2 Table
(Conditional means: Row totals fixed and column totals fixed)

When the OR =1, we have the **central hypergeometric model** just introduced.

\[
\text{Probability } [ \text{“a” given row and column totals given OR=1} ] = \frac{\binom{a + c}{a} \binom{b + d}{b}}{\binom{a + b + c + d}{a + b}}
\]

*(optional – for the interested reader):* When the OR ≠ 1, the correct probability model for the count “a” is a different hypergeometric distribution; this latter is called a non-central hypergeometric distribution. Interestingly, the probability calculation involves the magnitude of the OR which is now a number different from 1.

Probability [ “a” given row and column totals when OR ≠ 1] is the **non-central hypergeometric probability**

\[
= \frac{\binom{a + c}{a} \binom{b + d}{b} [OR]^y}{\sum_{u=\max(0,a-d)}^{\min(a+b,a+c)} \binom{a + c}{u} \binom{b + d}{a + b - u} [OR]^y}
\]

**Note – You will not need to work with the non-central hypergeometric distribution in this course.**
How to compute the p-value for this hypothesis test:

The solution for the p-value will be the sum of the probabilities of each possible value of the count “a” that are as extreme or more extreme relative to a null hypothesis odds ratio OR = 1.

**Step 1:**

If the row and column totals are fixed, what values of the count “a” are possible?

**Answer:** Because the column total is 4, the only possible values of the count “a” are 0, 1, 2, 3, and 4.

**Step 2:**

What are the probabilities of the five possible 2x2 tables?

**Answer:** We calculate the hypergeometric distribution likelihood for each of the 5 tables. While we’re at it, we’ll calculate the empirical odds ratio (OR) accompanying each possibility:

\[
\begin{array}{ccc|c|c}
\text{Disease} & \text{Healthy} & \text{Pr(table)} & \text{OR} \\
\hline
\text{Exposed} & 0 & 23 & 23 & .6875 & 0 \\
\text{Not exposed} & 4 & 232 & 236 & 255 & 259 \\
\hline
\text{a=0} & & & & \\
\hline
\text{Disease} & \text{Healthy} & \text{Pr(table)} & \text{OR} \\
\hline
\text{Exposed} & 1 & 22 & 23 & .2715 & 3.6 \\
\text{Not exposed} & 3 & 233 & 236 & 255 & 259 \\
\hline
\text{a=1} & & & & \\
\end{array}
\]
Step 3:

What is the p-value associated with the observed table?

Answer: The observed table has a count of “a”=2. Recalling our understanding of the ideas of statistical hypothesis testing, the p-value associated with this table is the likelihood of this table or one that is more “extreme”.

The more “extreme” tables are the ones with higher counts of “a” because these tables correspond to settings where the OR are as large or larger than the observed value of 11.1

\[
p-value = Pr[\text{Table with “a” } = 2] + Pr[\text{Table with “a” } = 3] + Pr[\text{Table with “a” } = 4] \\
= .0386 + .0023 + .0001 \\
= .0410
\]

Step 4:

Interpretation of Fisher’s Exact Test calculations.

Under the null hypothesis of no association between exposure and disease, the chances of obtaining an OR=11.1 or greater, when the row and column totals are fixed, are approximately 4 in 100.
## 7. Discrete Distributions - Themes

<table>
<thead>
<tr>
<th></th>
<th><strong>With Replacement</strong></th>
<th><strong>Without Replacement</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Framework</strong></td>
<td>A proportion $\pi$ are event</td>
<td>A fixed population of size $= N$ contains a subset of size $= M$ that are event</td>
</tr>
<tr>
<td><strong>Sampling</strong></td>
<td>Sample of size $= n$ with replacement</td>
<td>Sample of size $= n$ without replacement</td>
</tr>
<tr>
<td><strong>Outcome</strong></td>
<td># events $= x$</td>
<td># events $= x$</td>
</tr>
<tr>
<td><strong>Likelihood of outcome</strong></td>
<td>$\binom{n}{x} \pi^x (1 - \pi)^{n-x}$</td>
<td>$\binom{M}{x} \binom{N - M}{n - x} \binom{N}{n}$</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>Test of equality of proportion</td>
<td>Test of independence</td>
</tr>
</tbody>
</table>