Unit 3.
Discrete Distributions

“Chance favors only those who know how to court her”
- Charles Nicolle

In many research settings, the outcome variables are continuous (eg; height, weight, blood pressure, growth, blood lipid levels, etc) and their distributions are well described by a normal distribution model. We previously learned the importance of the normal distribution in the sampling distribution of the sample mean. Thus, the normal distribution is one of the most important distributions for us to understand.

There are also many research settings where the outcome variables are measured as counts (eg; number of deaths, number of events of nosocomial infections, number of lightening strikes, etc).

(1) We use the binomial distribution to model the chances of a given number of events in a known number of trials (eg – the probability of 5 nosocomial infections in a ward of 20 beds);

(2) We use the poisson distribution to model the chances of a given number of events when the event occurs rarely and we have no way of knowing how often the event did not happen (eg – the chances of 2 lightening strikes in Amherst, MA in 2016); and

(3) We use the (central) hypergeometric distribution to test the null hypothesis of no association in two-way tables of count data (eg – H₀: Exposure to alcohol (yes/no) is not associated with pancreatic cancer (yes/no)).

The binomial and poisson distributions are central to the modeling of count data.
# Table of Contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Objectives</td>
<td>3</td>
</tr>
<tr>
<td>1. Proportions and Rates in Epidemiological Research</td>
<td>4</td>
</tr>
<tr>
<td>2. Review - Bernoulli Distribution</td>
<td>9</td>
</tr>
<tr>
<td>3. Review - Binomial Distribution</td>
<td>12</td>
</tr>
<tr>
<td>4. Poisson Distribution</td>
<td>19</td>
</tr>
<tr>
<td>5. Hypergeometric Distribution</td>
<td>26</td>
</tr>
<tr>
<td>6. Fisher’s Exact Test of Association in a 2x2 Table</td>
<td>31</td>
</tr>
<tr>
<td>7. Discrete Distribution – Themes</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Appendices</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Discrete Distribution Calculators</td>
<td>37</td>
</tr>
<tr>
<td>1. Binomial Distribution</td>
<td>37</td>
</tr>
<tr>
<td>2. Poisson Distribution</td>
<td>39</td>
</tr>
<tr>
<td>3. Hypergeometric Distribution</td>
<td>41</td>
</tr>
<tr>
<td>B. Fisher’s Exact Test Calculators</td>
<td>43</td>
</tr>
<tr>
<td>C. Mean and Variance of Bernoulli((\pi))</td>
<td>45</td>
</tr>
<tr>
<td>D. The Binomial(n,(\pi)) is the Sum of n Independent Bernoulli((\pi))</td>
<td>46</td>
</tr>
<tr>
<td>E. The Poisson is an Extension of a Binomial</td>
<td>48</td>
</tr>
</tbody>
</table>
Learning Objectives

When you have finished this unit, you should be able to:

- Explain the distinction between continuous versus count data.
- Define and explain three probability distribution models for count data: *binomial*, *poisson*, and *hypergeometric*.
- Choose an appropriate probability model for a given set of count data: *binomial*, *poisson*, and *hypergeometric*.
- Calculate *binomial distribution* probabilities.
- Calculate *poisson probability* distribution probabilities.
- Perform and interpret *Fisher’s Exact Test* of association for data in a 2x2 table.
1. Proportions and Rates in Epidemiological Research

Counts are not assessed in isolation. For example, a count of 2 occurrences of cancer might have been observed among 100 exposed persons. Alternatively, a count of 2 occurrences of cancer might have been recorded by the CDC in its one year surveillance of a community exposed to water contamination.

Thus, often, we are analyzing proportions or rates. In epidemiologic research (as an example), the concepts of “proportion of” and “rate of” describe different aspects of disease occurrence.

**Proportion**

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.**

**Rate**

A rate is a count of event occurrence per unit of time. It is measured relative to an interval of time. It is not dimensionless, so be sure to report the denominator!

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

Valid range: 0 to ∞

E.g., - 100 persons are known to have smoked, collectively, for 1,000 pack years (Note – a “one pack year” unit of smoking corresponds to one pack a day, every day for 1 year). 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.

Some Commonly Used Proportions

**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>Denominator = # events that could have occurred</th>
<th># persons in the population at a point in time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator  = # events that actually occurred</td>
<td># persons having the disease at a point in time</td>
<td></td>
</tr>
<tr>
<td>Valid range:</td>
<td>0 to 1</td>
<td></td>
</tr>
<tr>
<td>Interpretation</td>
<td>The proportion of the population with disease at a point in time</td>
<td></td>
</tr>
<tr>
<td>Other names</td>
<td>Point prevalence</td>
<td></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 (“point 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.
Some Commonly Used Proportions - continued

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>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Denominator</strong> = # events that could have happened</td>
</tr>
<tr>
<td><strong>Numerator</strong> = # events that actually occurred</td>
</tr>
<tr>
<td><strong>Valid range:</strong></td>
</tr>
<tr>
<td><strong>Interpretation</strong></td>
</tr>
<tr>
<td><strong>Note !!</strong></td>
</tr>
</tbody>
</table>

Example:
In this example, the event of interest is the completion of a repeat screening mammogram, coded as 1=yes and 0=no.

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.
Some Commonly Used Rates

Incidence Density = \# disease onsets during interval of time
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>Denominator = # 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.(^a)</td>
</tr>
<tr>
<td></td>
<td>This sum is called “person time” and is (\sum_{i=1}^{N} (\text{time for } i^{th} \text{ person free of event}))</td>
</tr>
<tr>
<td></td>
<td>It is also called “person years” or “risk time”.</td>
</tr>
<tr>
<td>Numerator = # 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>

Tip! Be careful in the reporting of an incidence density. 
Don’t forget the scale of measurement of time. For example, here are 3 ways of saying the same thing: 
\[ I = 7 \text{ per “person year”} \]
\[ = \frac{7}{52} = 0.13 \text{ per “person week”} \]
\[ = \frac{7}{365.25} = 0.019 \text{ per “person day”} \]
What to Use?

Prevalence
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)

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

- Note - 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

Four, interrelated, probability distributions are considered in the analysis of counts, proportions, and rates.

1. Bernoulli; and
2. Binomial.
3. Poisson; and
4. Hypergeometric.

This section is a review of the Bernoulli Distribution. See also BIOSTATS 540, Unit 4. Bernoulli and Binomial, at:

http://people.umass.edu/biep540w/webpages/binomial.html

Example – The fair coin toss.

- We’ll use capital Z as our placeholder for the random variable name here:

  \[ Z = \text{Face of coin toss} \]

- We’ll use small \( z \) as our placeholder for a value of the random variable \( Z \):

  \[ \begin{align*}
  z &= 1 \text{ if “heads”} \\
  z &= 0 \text{ if “tails”}
  \end{align*} \]

- We’ll use \( \pi \) and \( (1-\pi) \) as our placeholder for the associated probabilities

  \[ \begin{align*}
  \pi &= \Pr[Z=1] \\
  (1-\pi) &= \Pr[Z=0]
  \end{align*} \]

  eg – This is the probability of “heads” and is equal to .5 when the coin is fair
Bernoulli Distribution ($\pi$) (also called “bernoulli trial”)
Equivalent to a Binomial(1, $\pi$)

A random variable $Z$ is said to have a Bernoulli Distribution if it takes on the value 1 with probability $\pi$ and takes on the value 0 with probability (1-$\pi$).

\[
\begin{array}{c|c}
\text{Value of } Z & P[Z = z] \\
\hline
1 & \pi \\
0 & (1 - \pi)
\end{array}
\]

This gives us the following expression for the likelihood of $Z=z$. \textbf{Tip} - Recall from BIOSTATS 540, Unit 5, that the likelihood function is called a probability density function and is written with the notation $f_Z(z)$. When the random variable is discrete (as is the case for the Bernoulli), we can write the following:

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

\[(1) \quad \mu = \text{Mean} = E[Z] = \text{Statistical Expectation of } Z
\]
\[\mu = \pi \quad \text{See page 44 for proof.}\]

\[(2) \quad \sigma^2 = \text{Variance} = \text{Var}[Z] = E[(Z-\mu)^2] = \text{Statistical Expectation of } (Z-\mu)^2
\]
\[\sigma^2 = \pi (1 - \pi) \quad \text{See page 44 for proof.}\]

A Bernoulli distribution is used to model the outcome of a SINGLE “event” trial

\textbf{Eg} – mortality, MI, etc.
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.

Later in this course (BIOSTATS 640 Unit 5. Logistic Regression), we’ll see that individual Bernoulli distributions are the basis of describing patterns of disease occurrence in a logistic regression analysis.

http://people.umass.edu/biep640w/webpages/logistic.htm
3. Review - Binomial Distribution

Recall from BIOSTATS 540 - When is the Binomial Distribution Used? -
The binomial distribution is used to answer questions of the form: “What is the probability that, in \( n \) independent success/failure trials with probability of success equal to \( \pi \), the result is \( x \) events of success?”

E.g. - the Binomial distribution model is used to calculate the chances of such things as
- 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?

What are \( n, \pi, \) and \( X \) in the Binomial Distribution?

\( n = \) number of independent trials (eg – the number of coin tosses performed, \( n=20 \))
\( \pi = \) Probability[individual trial yields “success” (eg – probability[single coin lands heads] = \( \frac{1}{2} \))
\( x = \) number of events of success that is obtained (eg – \( x=12 \) “heads”)

More generally:
What is the probability that \( n \) independent “event/non-event” trials, each with probability of event equal to \( \pi \) will yield \( x \) events?

<table>
<thead>
<tr>
<th>( Value \text{ of } X = )</th>
<th>( P[X = x] = )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( (1-\pi)^n )</td>
</tr>
<tr>
<td>1</td>
<td>( n \pi )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( x )</td>
<td>( \binom{n}{x} \pi^x (1-\pi)^{n-x} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( n )</td>
<td>( \pi^n )</td>
</tr>
</tbody>
</table>

Binomial Distribution (\( n, \pi \))

\( n \) independent Bernoulli Trials

A random variable \( X \) is said to have a Binomial (\( n,\pi \)) if it is the sum of \( n \) independent Bernoulli (\( \pi \)) trials.
f_X(x) = Likelihood | X = x | = \binom{n}{x} \pi^x (1-\pi)^{n-x} \text{ for } x=0, \ldots, n

(1) \mu = \text{Mean} = E[X] = \text{Statistical Expectation of } X
\quad \mu = n\pi

(2) \sigma^2 = \text{Variance} = \text{Var}[X] = E[(X-\mu)^2] = \text{Statistical Expectation of } (X-\mu)^2
\quad \sigma^2 = n\pi (1 - \pi)

**Binomial Probability Distribution Formula**

The **binomial formula** is the binomial distribution probability that you use to calculate a binomial probabilities of the form:

*What is the probability that, among n independent Bernoulli trials, each with probability of success = \pi, x events of “success” occur?*

The probability of obtaining exactly \(x\) events of success in \(n\) independent trials, each with the same probability of event success equal to \(\pi\):

\[
\Pr[X=x] = \binom{n}{x} \pi^x (1-\pi)^{n-x} = \frac{n!}{x! (n-x)!} \pi^x (1-\pi)^{n-x}
\]
Examples-

- What is the probability that 5 draws, with replacement, from an urn with 10 marbles (1 red, 9 green) will yield exactly 2 red? Answer: \( \text{Pr}[X=2] \) for Binomial \((n=5, \pi=1/10)\)

- What is the probability that among 100 vaccinated for flu, with subsequent probability of flu equal to .04, that 13 will suffer flu? Answer: \( \text{Pr}[X=13] \) for Binomial \((n=100, \pi=.04)\)

Tip! Here is a good on-line calculator for computing binomial probabilities:

http://faculty.vassar.edu/lowry/binomialX.html

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:**

http://faculty.vassar.edu/lowry/binomialX.html

“Event” is outcome of either “5” or “6”
\[ \text{Pr[event]} = \pi = .20 \]

\[ N = 20 \]

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

\[ \text{Pr}[X=3] = \binom{20}{3} [.20]^3 [1-.20]^{20-3} \]
\[ = \binom{20}{3} [.20]^3 [.80]^{17} \]
\[ = .2054 \]

Note – In the “k” box, enter “x”. In the “p” box, enter “π”

#2. **Solution:**

http://faculty.vassar.edu/lowry/binomialX.html

“Event” is outcome of either “7” or “8” or “9”
\[ \text{Pr[event]} = \pi = .30 \]

\[ N = 20 \]

\[ X \text{ is distributed Binomial}(N=20, \pi=.30) \]

Translation:
“At most 3 times” is the same as “less than or equal to 3 times”
\[ \text{Pr}[X \leq 3] = \text{Pr}[X=0] + \text{Pr}[X=1] + \text{Pr}[X=2] + \text{Pr}[X=3] \]
\[ = \sum_{x=0}^{3} \binom{20}{x} [.30]^x [.70]^{20-x} \]
\[ = \binom{20}{0} [.30]^0 [.70]^{20} + \binom{20}{1} [.30]^1 [.70]^{19} \]
\[ + \binom{20}{2} [.30]^2 [.70]^{18} + \binom{20}{3} [.30]^3 [.70]^{17} \]
=.10709
Review of the Normal Approximation for the Calculation of Binomial Probabilities

Use the Normal Approximation when:

\[(1) \, n \, \pi \geq 10; \text{ and} \]
\[(2) \, n \,(1 - \pi) \geq 10.\]

**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.

\[X \text{ distributed Binomial (N=180, } \pi=0.041) \text{ says that} \]
\[
\begin{align*}
\mu_{\text{BINOMIAL}} &= n\pi = (180)(.041) = 7.38 \\
\sigma^2_{\text{BINOMIAL}} &= n \pi (1 - \pi) = (180)(.041)(.959) = 7.08 \\
\sigma_{\text{BINOMIAL}} &= \sqrt{\sigma^2_{\text{BINOMIAL}}} = 2.66
\end{align*}
\]

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

\[
\begin{align*}
\text{Pr}[5 \leq X \leq 28] &= \text{Pr}[\frac{5-\mu}{\sigma} \leq \text{Normal}(0,1) \leq \frac{28-\mu}{\sigma}] \\
&= \text{Pr}[\frac{5 - 7.38}{2.66} \leq \text{Normal}(0,1) \leq \frac{28 - 7.38}{2.66}] \\
&= \text{Pr}[ -0.895 \leq \text{Normal}(0,1) \leq 7.752 ] \\
&\approx \text{Pr} [ -0.895 \leq \text{Normal}(0,1) ], \text{ because 7.752 is in the extreme right tail.} \\
&\approx \text{Pr} [ \text{Normal (0,1)} \leq +.895 ], \text{ because of symmetry of the tails of the normal} \\
&= .8146
\end{align*}
\]
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 with Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr [ X=k ]</td>
<td>Pr [ (k-1/2) ≤ X ≤ (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

When is the Poisson Distribution Used? -
The poisson distribution is used to answer questions pertaining to very rare events and are questions of the form: “What are the chances that in one year of surveillance in Amherst, MA, there will be 1 case of measles?

What is $\mu$ in the Poisson Distribution?

$\mu = \text{expected number of events in a specified frame}$ of observation. The frame might be

* time - a specified total time of surveillance (eg 1 year of surveillance)
* geographic area – a specified area on a map (eg –continental US)

E.g. – the poisson distribution model is used to calculate the chances of such things as
- What are the chances of 1 tornado strike hitting Amherst MA in 2016?
- What are the chances of the CDC recording 2 cases of ebola in the US in 2016?

Tip! Notice that we have a count of how many events did occur. But we do NOT have a count of how many events did not occur.

Tip! The Poisson distribution is a good choice for the modeling of count data that are rare.

Tip! In epidemiology, the Poisson distribution is used for the modeling of rates (as opposed to proportions)

Here, we will develop an understanding of the Poisson distribution using the idea of 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?

Note- We could just has easily use the idea of persons over space, instead of persons over time

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. See Appendix C for details.

- The concept of n persons (or n trials in a binomial) → 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!

### Poisson Distribution

If $X$ is distributed Poisson with mean $\mu$,

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

Expected value of $X$ is $E[X] = \mu$

Variance of $X$ is $Var[X] = \sigma^2 = \mu$  

That’s right – mean and variance are the same.

Where the mean $\mu$ is the parameter representing the product of the (incidence rate) $x$ (period of observation)

**Interpretation of $\mu$** – “This is the expected number of events we expect to get over this particular “frame” and this particular Poisson model”

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:
Suppose lung cancer occurs at a rate of 2 per 1000 pack years. Using a Poisson distribution model, calculate the probability of exactly 3 cases of lung cancer in 3000 pack years.

Solution:

Step 1: Solve for the Poisson mean parameter $\mu$ for the period of observation of interest.

$\mu = \text{(incidence rate)} \times \text{(period of observation)}$

$= (\frac{2}{1000 \text{ pack years}}) \times (3000 \text{ pack years})$

$= 0.002 \times 3000$

$= 6$

Thus, according to this Poisson model, we expect to see 6 cases of lung cancer over 3000 pack years.

Step 2: Identify desired calculation

“Calculate the probability of exactly 3 cases of lung cancer” is translated as follows

For $X$ distributed Poisson($\mu=6$),

Probability $[X=3] = ???$

If we expect 6, what are our chances of seeing 3 cases over 3000 pack years?

Step 3: Solve, by hand or with use of calculator on the web. Tip – If you are doing the calculation by hand using a hand calculator the exponentiation function button might appear as $\exp$ or it might appear as, simply, $e$.

$\text{Probability } [X=3] = \frac{\mu^x \exp(-\mu)}{x!} = \frac{6^3 \exp(-6)}{3!} = \frac{6^3 e^{-6}}{3!} = 0.0892$

http://stattrek.com/Tables/poisson.aspx
Example – continued.

This same example can be used to illustrate 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>(a) Binomial</th>
<th>(b) Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong> = # trials</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>π</strong> = Pr [ event ]</td>
<td></td>
<td>What happens as <strong>n</strong>→∞ and <strong>π</strong>→0?</td>
</tr>
<tr>
<td><strong>Expected # events</strong></td>
<td><strong>nπ</strong></td>
<td><strong>μ</strong></td>
</tr>
<tr>
<td><strong>Pr [ <strong>X</strong>=x events ]</strong></td>
<td>( \binom{n}{x} \pi^x (1-\pi)^{n-x} )</td>
<td>( \frac{\mu^x \exp(-\mu)}{x!} )</td>
</tr>
</tbody>
</table>

where \( \exp = \) numerical constant = 2.718

**Solution using the Binomial:**

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

**Solution using the Poisson:**

\[
\mu = \lambda T = (.002/\text{pack year})(3000 \text{ pack years}) = 6
\]

\[
\text{Pr}[3 \text{ cases cancer}] = \frac{6^3 \exp[-6]}{3!} = .0892
\]

Pretty close ….
An example of the Poisson sampling design is a cross-sectional study

<table>
<thead>
<tr>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>a+c</td>
<td>b+d</td>
</tr>
<tr>
<td>a+b</td>
<td>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]

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

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) $\Pr[\text{exposure given disease}] \neq \Pr[\text{exposure given no disease}]$

$$\frac{\mu_{11}}{\mu_{11} + \mu_{21}} \neq \frac{\mu_{12}}{\mu_{12} + \mu_{22}}$$
An example of the Binomial sampling design is a cohort 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>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

a = # exposed persons develop disease
b = # exposed persons who do not develop disease
c = # unexposed persons who develop disease
d = # unexposed persons who do not develop disease

- 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]

$$\pi_1 \neq \pi_2$$
Another example of the Binomial sampling design is a case-control 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>a+c fixed</td>
<td>a+b</td>
<td>c+d</td>
</tr>
<tr>
<td>b+d fixed</td>
<td>a+b+c+d</td>
<td></td>
</tr>
</tbody>
</table>

- a = # persons with disease whose recall reveals exposure
- c = # persons with disease whose recall reveals no exposure
- b = # healthy persons whose recall reveals exposure
- d = # healthy persons whose recall reveals no exposure

- The counts a and b are each free to vary and follow their own distribution.
- The counts c and d do not vary because of the fixed column totals
- a and b are 2 independent Binomial random variables.
- Let’s denote the means of these $\theta_1$ and $\theta_2$.

Likelihood[2x2 table]

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

An association between exposure and disease means the following.
Pr[exposure among disease] ≠ Pr [exposure among healthy]

$$\theta_1 \neq \theta_2$$
5. Hypergeometric Distribution

The (central) hypergeometric distribution is the null hypothesis probability model in a Fisher’s Exact test of the null hypothesis of no association. Consider the following 2x2 table example.

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.

<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>

Some things to notice here are the following and they pertain to the null hypothesis of “no association”:

1. Overall, among the 259 pregnancies, 4 ended in spontaneous abortion. This is an overall rate of 4/259 = 0.0154 or 1.5%

2. Assumption of the null hypothesis model says that the overall 1.5% rate also applies to: (i) the video display terminal group and (ii) the non-video display terminal group.

3. Video Display Terminal Group: Application of null hypothesis 1.5% rate →
   Among the 23 women working with video display terminals (these are our “exposed”),
   Null expected number of spontaneous abortions = (.0154)*(#women) = (.0154)(23) = .3552

4. Non Video Display Terminal Group: Application of null hypothesis 1.5% rate →
   Among the 236 women who did NOT work with video display terminals,
   Null expected number of spontaneous abortions = (.0154)*(#women) = (.0154)(236) = 3.64

5. Comparison of null hypothesis expected counts with observed counts:
   The null hypothesis expected counts were .3552 and 3.64 events, respectively.
   The actual observed counts were 2 and 2 events, respectively
   The exposed women experienced a worrisomely high number of spontaneous abortions: 2 instead of 0.3552. Is this discrepancy statistically significant?
Under the null hypothesis of “no association”, what were the chances that 2 of the 4 events of spontaneous abortions occurred among the exposed?

The solution below is a “hand calculation” that makes use of a calculator on the internet for the solution of combinations and permutations. The solution can also be obtained by using a hypergeometric distribution calculator. See Appendix A.3 – Hypergeometric Distribution.

- If 2 abortions occurred among the VDT workers, that means that:
  2 abortions occurred among the VDT workers and 2 abortions occurred in NON-VDT workers.
  This is because there were 4 events of abortion in total.

- Under “no association”, overall:
  \# ways to choose 4 abortions from 259 pregnancies is \( \binom{259}{4} = 183,181,376 \)

- Under “no association”, among the 23 “exposed”:
  \# ways to choose 2 abortions from 23 pregnancies is \( \binom{23}{2} = 253 \)

- Under “no association”, among the 236 “NON-exposed”:
  \# ways to choose 2 abortions from 236 pregnancies is \( \binom{236}{2} = 27,730 \)

- Thus, under the null hypothesis of “no association” the probability that 2 of the 4 abortions occurred among the 23 exposed by chance is the hypergeometric probability

\[
\frac{23 \choose 2 \cdot 236 \choose 2}{259 \choose 4} = \frac{253 \cdot 27,730}{183,181,376} = .038
\]

http://www.mathsisfun.com/combinatorics/combinations-permutations-calculator.html
Refresher – In a 5 card hand of poker, what are the chances of obtaining a hand with exactly 2 queens?

Answer: .0399, representing a 4% chance, approximately.

- 2 queens in a 5 card hand means: 2 queens and 3 NON queens.
- # ways to obtain a 5 card hand from a 52 card desk is \( \binom{52}{5} = 2,598,960 \)
- # ways to obtain 2 queens from 4 possible queens is \( \binom{4}{2} = 6 \)
- # ways to obtain 3 non-queens from 48 possible non-queens is \( \binom{48}{3} = 17,296 \)
- \( \to \) # 5 card hand with 2 queens and 3 non-queens is \( \binom{4}{2}\binom{48}{3} = (6)(17,296) = 103,776 \)
- \( \text{Pr}[2 \text{ queens in five card hand}] = \frac{\text{(# of 5 card hands with 2 queens & 3 non-queens)}}{\text{( # 5 card hands total) }} \)
  \[ = \frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} = \frac{(6)(17,296)}{2,598,960} = .03993 \]

http://www.mathsisfun.com/combinatorics/combinations-permutations-calculator.html
The Null Hypothesis “No Association” model of a 2x2 tables of counts is the Central Hypergeometric. Often, we are interested in a test of “no association” in the setting of a 2x2 table analysis of exposure-disease count data. In the table below, we are interested in the count “a” because this is the number of cases among the exposed. The question is: Is the observed count “a” statistically significantly different from what we expected under the null hypothesis of no association?

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</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>(a+c) fixed</td>
<td>(b+d) fixed</td>
<td></td>
</tr>
</tbody>
</table>

(a+b) fixed
(c+d) fixed
n = a + b + c + d

- If “a” cases occurred among the exposed, that means that: “a” cases occurred among the exposed and “c” cases occurred among the non-exposed. This is because there are (a+c) cases in total.

- Under the null hypothesis model of “no association”, overall:
  
  \[
  \text{# ways to choose (a+c) cases from the total is } \binom{n}{a+c} = \binom{a+b+c+d}{a+c}
  \]

- Under the null hypothesis model of “no association”, among the (a+b) “exposed”:
  
  \[
  \text{# ways to choose “a” cases from “(a+b)” exposed is } \binom{a+b}{a}
  \]

- Under the null hypothesis model of “no association”, among the (c+d) “NON-exposed”:
  
  \[
  \text{# ways to choose “c” cases from the “(c+d)” NON-exposed is } \binom{c+d}{c}
  \]

- Thus, under the null hypothesis of “no association” the probability that “a” of the “(a+c)” cases occurred among the exposed is

\[
\binom{a+b}{a} \binom{c+d}{c} = \binom{a+b}{a} \binom{c+d}{c} = \binom{n}{a+c} \binom{a+b+c+d}{a+c}
\]
Natural Result – It works for both cohort (prospective) and case-control (retrospective) 2x2 tables.

You might have noticed that we did the calculation for a 2x2 table arising from a cohort study design, where the number of exposed (a+b) and the number of non-exposed (c+d) are fixed. We would have gotten the same result if we had done the calculation for a 2x2 table arising from a case-control study design where the number of cases (a+c) and the number of controls (b+d) are fixed.

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 null hypothesis “no association” (chance) model for the probability of observing counts “a”, “b”, “c” and “d” in a 2x2 table. It is the central hypergeometric probability distribution 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>
<tr>
<td></td>
<td></td>
<td>259</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 Binomial 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?

Recall from page 26 that we calculated the expected number of events of spontaneous abortions among the 23 exposed under the null hypothesis model assumption of “no association”. We expected to see 0.3552 events. We observed, instead, 2 events. Is the observed 2 statistically significantly greater than the null hypothesis expected 0.3552? We will use the central hypergeometric distribution to solve for a p-value.
What is the null hypothesis model probability (likelihood) of the configuration of counts “a”, “b”, “c”, and “d” if the row and column totals are held fixed?

Recall the layout of our 2x2 table. If we hold the row and column totals as fixed, then the probability we are after is an example of what is called a conditional probability. A look below reveals that, once the row and column totals are held fixed, only one cell count is free to vary. This is because every other cell count can then be obtained by subtraction. Let the free to vary cell count be the count “a”. Thus we are defining a conditional probability model for the count “a”, conditioning on the row and column totals.

<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>b</td>
</tr>
<tr>
<td></td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

**Conditional Probability Distribution of “a” in the 2x2 Table**

(Row totals fixed and column totals fixed)

**Null Hypothesis Model** (We used this for the calculation of p-values)

Scenario 1 (Null is true): “No Association (Odds Ratio = 1)” Model
Central hypergeometric model.

\[
\text{Probability [ # with exposure and with disease = a] } = \frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{a+b+c+d}{a+b}}
\]

**Alternative Hypothesis Model** (We use this for sample size and power calculations – not discussed here)

Scenario 2 (Alternative is true): “Association (Odds Ratio ≠ 1)” Model
NON - Central hypergeometric model.

\[
\text{Probability [ # with exposure and with disease = a] } = \sum_{u=\max(b,d)}^{\min(a+c,a+b)} \frac{\binom{a+c}{a+u} \binom{b+d}{b+u}}{\binom{a+b+c+d}{a+b-u}} [OR]^u
\]
Example - continued:
What is the statistical significance (p-value) of 2 abortions under the null hypothesis of “no association”?

Answer:

\[ p\text{-value} = \text{Probability} \{ 2 \text{ abortions among the exposed or more extreme | null model is true} \} \]

\[ = \text{Probability} \{ 2 \text{ abortions among the exposed | null true} \} \\
+ \text{Probability} \{ 3 \text{ abortions among the exposed | null true} \} \\
+ \text{Probability} \{ 4 \text{ abortions among the exposed | null true} \} \]

Illustration: For illustration purposes, let’s calculate the null hypothesis central hypergeometric distribution probabilities for all of the 5 tables that are possible once we hold fixed the row and column totals. While we’re at it, we’ll calculate the empirical odds ratio (OR) accompanying each possibility. I found two nice online calculators to help us out:

(1) Calculation of null hypothesis “no association” central hypergeometric probabilities:
http://stattrek.com/online-calculator/hypergeometric.aspx

(2) Calculation of odds ratios and associated 95% confidence intervals:

\[
\begin{array}{l|cc|c|c}
\text{Disease} & \text{Exposed} & \text{Not exposed} & \text{Healthy} & \text{Pr(table)} \\
\hline
\text{Exposed} & a = 0 & 23 & 23 & .6875 \\
\text{Not exposed} & 4 & 232 & 236 & \text{OR=0} \\
\text{ } & 4 & 255 & 259 & \\
\hline
\end{array}
\]

\[
\begin{array}{l|cc|c|c}
\text{Disease} & \text{Exposed} & \text{Not exposed} & \text{Healthy} & \text{Pr(table)} \\
\hline
\text{Exposed} & a = 1 & 22 & 23 & .2715 \\
\text{Not exposed} & 3 & 233 & 236 & \text{OR=3.6} \\
\text{ } & 4 & 255 & 259 & \\
\hline
\end{array}
\]
a=2  

Note – This is our observed …

<table>
<thead>
<tr>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a = 2</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Not exposed</td>
<td>4</td>
</tr>
</tbody>
</table>

a=3

<table>
<thead>
<tr>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a = 3</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Not exposed</td>
<td>4</td>
</tr>
</tbody>
</table>

a=4

<table>
<thead>
<tr>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a = 4</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Not exposed</td>
<td>4</td>
</tr>
</tbody>
</table>

Thus,

\[
p-value = \text{Probability } [2 \text{ abortions among the exposed } \text{ or more extreme}]
\]

\[
= \text{Probability } [2 \text{ abortions among the exposed } | \text{ null true}] + \text{Probability } [3 \text{ abortions among the exposed } | \text{ null true}] + \text{Probability } [4 \text{ abortions among the exposed } | \text{ null true}]
\]

\[
= .0386 + .0023 + .0001
\]

\[
= .0410
\]
Interpretation of Fisher’s Exact Test calculations.

The assumption of the null hypothesis of “no association”, applied to the data in this 2x2 table has led to a reasonably unlikely outcome (p-value = .04), suggesting statistical rejection of the null hypothesis. We conclude that these data provide statistically significant evidence of an association of exposure to video display terminals in the workplace during pregnancy with change in risk of spontaneous abortion.

Stata Illustration

<table>
<thead>
<tr>
<th></th>
<th>col</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>-----</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>row</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>21</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>234</td>
<td>236</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>255</td>
<td>259</td>
<td></td>
</tr>
</tbody>
</table>

Fisher’s exact = 0.041
1-sided Fisher’s exact = 0.041

Nature Population/Sample Observation/Data Relationships/Modeling Analysis/Synthesis
## 7. Discrete Distributions - Themes

<table>
<thead>
<tr>
<th>Framework</th>
<th>With Replacement</th>
<th>Without Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A proportion</strong> $\pi$ of the outcomes are events</td>
<td>A fixed population of size $=N$ contains a subset of size $=M$ that are events</td>
<td></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>

### Nature

Population/ Sample

### Observation/

Data

### Relationships/

Modeling

### Analysis/

Synthesis
Appendix A
Discrete Distribution Calculators

1. Binomial Distribution

<table>
<thead>
<tr>
<th>Nature</th>
<th>Population/ Sample</th>
<th>Observation/ Data</th>
<th>Relationships/ Modeling</th>
<th>Analysis/ Synthesis</th>
</tr>
</thead>
</table>

Online Calculator

http://www.vassarstats.net/binomialX.html

<table>
<thead>
<tr>
<th>n</th>
<th>k</th>
<th>p</th>
<th>q</th>
<th>Calculate</th>
<th>Reset</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
<td>.67</td>
<td>0.32999999</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Enter
n = # trials
k = # events (we called this x)
p = probability (we called this \( \pi \))

Click calculate

Example –
In a binomial experiment with n=6 trials, success probability \( \pi = .67 \) and \( X= \)number of successes ....

Pr \( [X = 2] \) = .0798
Pr \( [X \leq 2] \) = .0968
Pr \( [X \geq 2] \) = .98298

etc … scroll down for more.
Stata Illustration using probcalc

Before you begin – Download the function probcalc
1. Launch stata
2. In the command window, type: ssc install probcalc
3. Recommended: In the command window, type: help probcalc

Scroll down until you see the help information for the Binomial distribution. You should see:

Commands for the Binomial Distribution:

- `probcalc b #n #p exactly #x` (Note: bold represent
  The probability distribution based on n and p, densiti

- `probcalc b #n #p dist`
  Probability of observing at most x events, \( P(X=x) \)

- `probcalc b #n #p atmost #x`
  Probability of observing at least x events, \( P(X\geq x) \)

- `probcalc b #n #p atleast #x`

Example –
In a binomial experiment with \( n=6 \) trials, success probability \( \pi = .67 \), what is the
probability of exactly \( X=2 \) successes? Answer: .07985

. * Illustration of binomial calculation using probcalc

. * For Binomial with \( n=6 \) and success probability=.67, calculate \( P[X=2] \)

. probcalc b 6 .67 exactly 2
Distribution: Binomial
n=6
p=.67
option:exactly
x=2
\( P(X=2) = .07985399 \)
2. Poisson Distribution

**Online Calculator**

http://stattrek.com/online-calculator/poisson.aspx

Enter values for the first 2 rows
- $x = \text{outcome} = \# \text{ events}$
- $\mu = \text{expected number of events over period of interest}$

Click **calculate**

**Example (page 21 of notes)**
For a Poisson process with $\mu = 6$, what is the probability of exactly 3 events?

Pr $[X = 3] = 0.08923$

*etc … scroll down for more.*
Stata Illustration using `probcalc`

If you have not already done so ... download the function `probcalc`
1. Launch stata
2. In the command window, type: `ssc install probcalc`
3. Recommended: In the command window, type: `help probcalc`
   Scroll down until you see the help information for the Poisson distribution. You should see:

Commands for the Poisson Distribution:

- Probability of observing exactly \( x \) events, \( P(X=x) \)
  
  `probcalc p #mu exactly #x`

  The probability distribution based on \( \mu \), densities \( \epsilon \)
  
  `probcalc p #mu dist`

  Probability of observing at most \( x \) events, \( P(X<=x) \)
  
  `probcalc p #mu atmost #x`

  Probability of observing at least \( x \) events, \( P(X>=x) \)
  
  `probcalc p #mu atleast #x`

Example –
In a Poisson process with expected number of events \( \mu = 6 \), what is the probability of exactly \( X=3 \) events? Answer: \( .08923 \)

```
. * For Poisson with expected number of events \( \mu=6 \), calculate \( \Pr[X=3] \)

. probcalc p 6 exactly 3

Distribution: Poisson
\( \mu=6 \)
option: exactly
\( x=3 \)
\( P(X=3) = .08923508 \)
```
3. (Central) Hypergeometric Distribution

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

23 = Sample Size

4 = Population size

# successes in population = 4

Online Calculator

http://stattrek.com/online-calculator/hypergeometric.aspx

Enter values for the first 4 rows

Population Size = n = total of 2x2 table
# Successes in Pop = (a+c) = column 1 total
Sample Size = (a+b) = row 1 total
# successes in sample = count in “a”

Click calculate

Example (page 28 of notes)

Under the null hypothesis of “no association”, what is the probability of 2 abortions among 23 exposed and 2 abortions among 236 NON-exposed?

Pr [ X = 2 ] = 0.038299
Stata Illustration using `display hypergeometricp()`

This command is included in Stata already – no need to download!

Recommended: In the command window, type: `help hypergeometricp`

*You should see…*

```
`probability_mass_distribution_function`

`hypergeometricp(N, K, n, k)`

Domain `N`: 2 to 1e+5
Domain `K`: 1 to `N-1`
Domain `n`: 1 to `N-1`
Domain `k`: `max(0,n-M*K)` to `min(K,n)`

Range: 0 to 1 (right exclusive)

Description: returns the hypergeometric probability of `k` successes (where success is obtaining an element with the attribute of interest) out of a sample of size `n`, from a population of size `N` containing `K` elements that have the attribute of interest.
```

**Example –**

For a 2x2 table with total `N=259`, `M=4` events total and sample of size `n=23`, what is the “no association” probability that the sample includes `X=2` successes?  **Answer: .03829914**

```
* Central Hypergeometric with `N=259`, # Events in Population=`4`, Sample Size=`23` and `X=2`

display hypergeometricp(259, 4, 23, 2)
.03829914
```
Appendix B

Fisher Exact Test Calculators

Online Calculator
http://www.quantitativeskills.com/sisa/statistics/fisher.htm

Enter values for

a = row 1, col 1
b = row 1, col 2
c = row 2, col 1
d = row 2, col 2

Click calculate

Example (page 34 of notes)
scroll down......

p-value = Pr \{ X \geq 2 \} = .0406
Stata Illustration using the “immediate” command `tabi` with option `exact`
This command is also included in Stata already.
Recommended: In the command window, type: `help tabi`

```
Immediate form
   . tabi 30 18 \ 18 14

Immediate form, 2 x 3 table
   . tabi 30 18 38 \ 13 7 22

Add Fisher's exact test
   . tabi 30 18 38 \ 13 7 22, ch2 exact

3 by 2 table, all measures of association
   . tabi 30 13 \ 18 7 \ 38 22, all exact
```

```
aved results

Example –
For a 2x2 table with total N=259, M=4 events total and sample of size n=23, what is the p-value for the Fisher Exact test of “no association”? Answer: .03829914
```

```
* Fisher Exact test for 2x2 table
   . tabi 2 21\2 234, exact

<table>
<thead>
<tr>
<th>row</th>
<th>col</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>Total</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>234</td>
<td>236</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>255</td>
<td>259</td>
</tr>
</tbody>
</table>

Fisher's exact = 0.041
1-sided Fisher's exact = 0.041
```
Appendix C
Mean (μ) and Variance (σ²) of a Bernoulli Distribution

Mean of Z = μ = π

The mean of Z is represented as E[Z]. recall: “E” stands for “expected value of”

\[ E[Z] = \pi \]

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 ].

Var[Z] = π(1-π) because the following is true:

\[
\text{Var}[Z] = E [ (Z-\pi)^2 ] = \sum_{\text{All possible } z} (z-\pi)^2 \text{ Probability}[Z=z] \\
= [(0-\pi)^2]\Pr[Z=0]+[(1-\pi)^2]\Pr[Z=1] \\
= [\pi^2] (1-\pi) + [(1-\pi)^2] (\pi) \\
= \pi (1-\pi) [ \pi + (1-\pi) ] \\
= \pi (1-\pi)
\]
Appendix D

The Binomial(n, π) is the Sum of n Independent Bernoulli(π)

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 (π) with Z=1 when the event occurs and Z=0 when it does not.

Tip - The use of the “0” and “1” coding scheme, with “1” being the designation for event occurrence, is key, here.

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

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

Consider what happens if we 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.

X represents the net number of successes in a set of independent Bernoulli trials. A simple example is the outcome of several coin tosses (e.g., “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 arrangements 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.

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

\( x \) events \hspace{2cm} \( N-x \) non events

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

**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.

\[
Pr[X = x] = Pr \left[ \sum_{i=1}^{n} Z_i = x \right] = \binom{n}{x} \pi^x (1-\pi)^{n-x}
\]
Appendix E
The Poisson Distribution is an Extension of Binomial

- 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 begins with a Binomial Distribution Situation.

- We begin by constructing a binomial likelihood situation. Let

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

- 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.

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

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

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 $\lambda T = \mu$.

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

$$= \binom{n}{x} \left(\frac{\lambda T}{n}\right)^x \left[1 - \frac{\lambda T}{n}\right]^{n-x}$$
Work with each term on the right hand side one at a time:

- 1st term -

\[
\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n(n-1)(n-2)...(n-x+1)}{x!(n-x)!} = \frac{n(n-1)(n-2)...(n-x+1)}{x!}
\]

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

Thus, as \( n \rightarrow \infty \)

\[
\binom{n}{x} \rightarrow \frac{n^x}{x!}
\]

- 2nd term -

\[
\left[ \frac{\lambda T}{n} \right]^x = \left[ \frac{\lambda}{n} \right]^x \left[ \frac{1}{n} \right]^x = \left[ \mu \right]^y \left[ \frac{1}{n} \right]^x
\]

Thus,

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

- 3rd 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 \( n \rightarrow \infty \)

\[
\left[ 1 - \frac{\lambda T}{n} \right]^n = \left[ 1 - \frac{\mu}{n} \right]^n \rightarrow e^{-\mu}
\]

Thus, as \( n \rightarrow \infty \)

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

Thus, as \( n \rightarrow \infty \)

\[
\left[ 1 - \frac{\lambda T}{n} \right]^x \rightarrow 1
\]

Now put together the product of the 4 terms and what happens as \( n \rightarrow \infty \)

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

as \( n \rightarrow \infty \)

\[
\frac{n^x}{x!} \left\{ [\mu]^x \left[ \frac{1}{n} \right]^x \right\} \left\{ e^{-\mu} \right\} \{ 1 \} = \frac{\mu^x e^{-\mu}}{x!}
\]

\[
= \frac{\mu^x e^{-\mu}}{x!} = \frac{\mu^x \exp(-\mu)}{x!}
\]
Poisson Distribution

If X is distributed Poisson (µ),

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

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

Variance of X is \( Var[X] = \sigma^2 = \mu \)

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