Unit 2.
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*. The normal distribution is one of the most important distributions for us to understand.

Moving on! There are also many research settings where the outcome variables are measured as *counts* (e.g., number of deaths, number of events of nosocomial infections, number of lightning strikes, etc.) **and other models are needed**.

(1) We use the *binomial* distribution to model the chances of a given number of events in a known number of trials (e.g., 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 (e.g., the chances of 2 lightning strikes in Amherst, MA in 2024); and

(3) We use the *(central) hypergeometric* distribution **a lot in the analysis of epidemiological tables; e.g.,** to test the null hypothesis of no association in **a** two-way table of count data (e.g., H_O: 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. The *central hypergeometric* distribution is the null hypothesis model of the Fisher Exact Test.

Table of Contents

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 distribution* probabilities.
- Perform and interpret *Fisher's Exact Test* of general association for data in a 2x2 table.

1. Proportions and Rates in Epidemiological Research

Counts are not assessed in isolation. For example, consider a count of 2 occurrences of cancer. Were these 2 cases that occurred among 100 exposed persons? Or were these 2 occurrences of cancer recorded by the CDC in a one-year surveillance (say, of a community exposed to water contamination). A count of 2 in isolation is uninterpretable. We need a "relative to something", a denominator!.

Thus, often, we analyze ***proportions*** or ***rates***. In epidemiology, 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" = 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 units of time that are being used!**

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

Valid range: 0 to ∞

Example: Suppose 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 = 3/1000 pack years

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

Incidence densities are examples of rates.

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

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 (or space).

Some Commonly Used Proportions

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

Prevalence	For example ….
Denominator = # events that could have occurred	# persons in the population at a point in time
Numerator = # events that actually occurred	# persons having the disease at a point in time
Valid range:	0 to 1
Interpretation	The proportion of the population with disease at a point in time
Other names	Point prevalence

Example:
Suppose that in 2018, suppose 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 the 2018 registry cohort was

$$P = \left[\frac{528}{16{,}529}\right]100\% = \left[0.0319\right]100\% = 3.19\%$$

These 528 women might be excluded from the analyses to determine the factors associated with repeat screening mammogram.

Some Commonly Used Proportions - *continued*

Cumulative incidence = $\frac{\text{\# new disease onsets during interval of time}}{\text{\# persons at risk in population at start of interval}}$

Cumulative Incidence	For example ….
Denominator = # events that could have happened	# 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.
Numerator = # events that actually occurred	# persons developing the disease during the time interval.
Valid range:	0 to 1
Interpretation	The proportion of healthy persons who go on to develop disease over a specified time period.
Note!	**Key -** We assume that every person in the population at the start was followed for the entire interval of time.

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

As of January 1, 2018, there were 9,485 women in the New York State Breast and Cervical Cancer Screening Program registry with a negative mammogram, and for whom: 1) there was no history of breast cancer; and 2) complete data were available.

2,604 obtained a repeat screening mammogram during the 6-year period January 1, 2018 – December 31, 2023.

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

$$CI_{6\text{-year}} = \left[\frac{2{,}604}{9{,}485}\right]100\% = \left[0.2745\right]100\% = 27\%$$

We might perform additional analyses focused on the identification of its correlates.

Some Commonly Used Rates

Incidence Density = $\frac{\text{\# new disease onsets during over total time at risk}}{\text{Total time at risk = sum of individual lengths of time at risk}}$

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

Tip! Be careful in the reporting of an incidence density.

Don't forget to include in your reporting *the span of surveillance (here - time frame) and the units.* For example, here are 3 ways of saying the same thing:

7 per "person year" is the same as

7/52 = 0.13 per "person week" which is also the same as

7/365.25 = 0.019 per "person day" and so on ...

What to Use?

The goal is to describe occurence of disease. We have choices - prevalence, cumulative incidence, incidence density.

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/causal studies.

- Susceptibility and duration of disease contribute to prevalence which means that
- prevalence = function(susceptibility, incidence, survival), making it a challenge to interpret

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.

Challenge - 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

This is an introduction to four (4) probability distributions that 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. **For a more detailed introduciton, s**ee again, *BIOSTATS 540, Unit 6. Bernoulli and Binomial*, at:

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

Example – The fair coin toss.

- By convention, we use **capital Z** as our placeholder for a random variable that is distributed Bernoulli.
  (Note - Sorry, this is a different Z than the Z introduced as a Z-score in Unit 1, Review of BIOSTATS 540):

  Z = Face of coin toss

- We'll use **small z** as our placeholder for a value of the random variable Z:
  z = 1 if "heads"
  z = 0 if "tails"

- We'll use **π and (1-π)** as our placeholder for the associated probabilities
  π = Pr[Z=1] **e.g., this is the probability of "heads" and is equal to .5 when the coin is fair**
  (1-π) = Pr[Z=0]

Bernoulli Distribution (π) (also called "Bernoulli trial")

Equivalent to a Binomial(1, π)

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

Value of Z =	P[Z = z] =
1	π
0	$(1 - \pi)$

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

Review - Recall from BIOSTATS 540, 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:

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

(1) μ = Mean = E[Z] = Statistical Expectation of Z

$\mu = \pi$ *See BIOSTATS 540 Unit 6, page 13 for proof.*

(2) σ^2 = Variance = Var[Z] = E[$(Z-\mu)^2$] = Statistical Expectation of $(Z-\mu)^2$

$\sigma^2 = \pi\,(1 - \pi)$ *See BIOSTATS 540 Unit 6, page 13 for proof.*

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

e.g., mortality, MI, etc.

In epidemiology, the Bernoulli probability distribution is used to model a "yes/no" ("event"/"non-event") outcome of event in 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 π; or
2) "non-event" with probability $(1-\pi)$

The Bernoulli probability model distribution **associates the two possible states with their associated probabilities**.

We have what we need to define the Bernoulli probability distribution **for Z ~ Bernoulli (π)**:

Outcome/State, z	**Likelihood, Pr [Z = z]**
1 ("event")	$\Pr [Z = 1] = \pi$
0 ("Non-event")	**Pr [Z = 0] =** $(1 - \pi)$

Check it out! Since the value of Z can only be z=1 or z=0, we can exploit this to write a single formula**, called the** probability density function **denoted** $f_Z(z)$**,** for computing likelihoods:

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

Later in this course (*BIOSTATS 640 Unit7. Logistic Regression*), we'll see that individual Bernoulli probability distribution models 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

An extension of Bernoulli distribution considers more than 1 trial and is called the Binomial distribution. It is used to model the chances of such things as

- What is the probability that 2 of 6 graduate students are first generation graduate students?
- What is the probability that of 100 infected persons, 4 will die within a year?

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, each **of which is a Bernoulli trial** with the same probability of success equal to π, the result is x events of success?"

What are n, π, and X in the Binomial Distribution?

n = number of independent trials (e.g., the number of coin tosses performed, n=20)
π = Probability[individual trial yields "success" (e.g., probability[single coin lands heads] = ½)
x = number of events of success that is obtained (e.g., x=12 "heads")

More generally:

What is the probability that n independent "event/non-event" trials, each with probability of event equal to π will yield x events?

X ~ Binomial Distribution (n, π)
X = # events in independent Bernoulli Trials

A random variable X is said to be distributed Binomial (n,π) if it is the sum of n independent Bernoulli (π) trials.

Value of X =	Probability X=x is Pr[X = x] =
0	$(1-\pi)^n$
1	$n\,\pi\,(1-\pi)^{n-1}$
…	…
x	$\binom{n}{x}\pi^x(1-\pi)^{n-x}$
…	…
n	π^n

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

$f_X(x)$ = Likelihood [X = x] = $\binom{n}{x}$ π^x $(1-\pi)^{n-x}$ for x=0, …, n

(1) μ = Mean = E[X] = Statistical Expectation of X
$\mu = n\pi$

(2) σ^2 = Variance = Var[X] = E[$(X-\mu)^2$] = Statistical Expectation of $(X-\mu)^2$
$\sigma^2 = n\pi\ (1-\pi)$

Binomial Probability Distribution
Formula for Calculating Probabilities

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

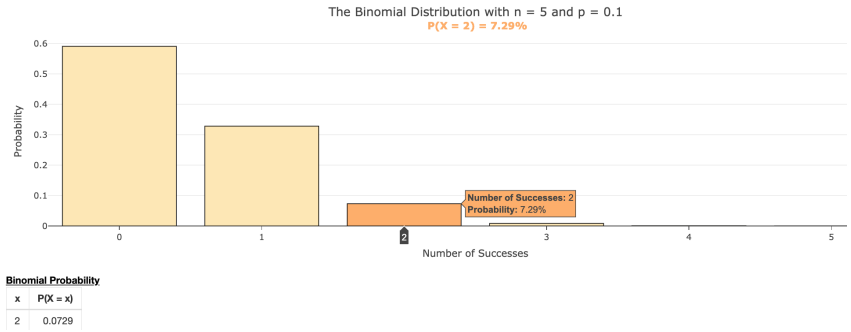
What is the probability that, among n independent Bernoulli trials, each with probability of success = π, x events of "success" occur?

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

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

Recall.
n! = (n)(n-1)(n-2)....(3)(2)(1)

Example - 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: Pr[X=2] for Binomial (n=5, π=1/10) **= .0729**

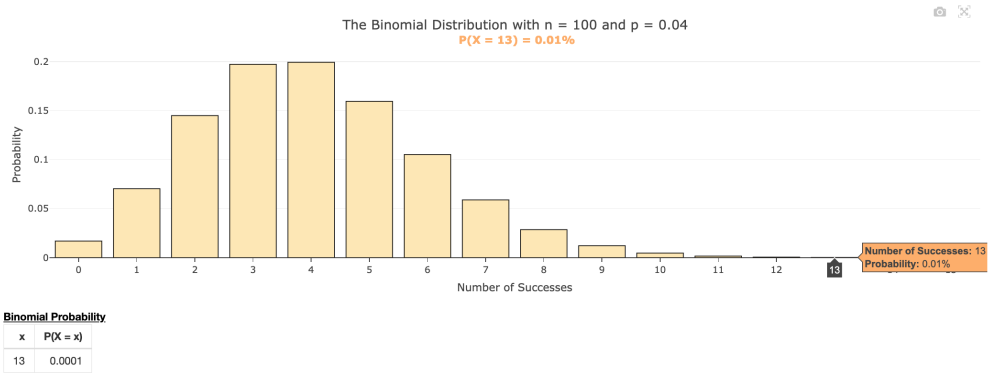


https://istats.shinyapps.io/BinomialDist/

R Solution

```
> # Binomial(n=5, p=.10)  Solution for Pr[X=2]
> dbinom(x=2, size=5, prob=.10)
[1] 0.0729
```

Example - What is the probability that among 100 vaccinated for flu, with subsequent probability of flu equal to .04, that 13 will suffer flu? Answer: Pr[X=13] for Binomial (n=100, π=.04**) = .0001**



https://istats.shinyapps.io/BinomialDist/

R Solution

```
> # Binomial(n=100, p=.04)  Solution for Pr[X=13]
> dbinom(x=13, size=100, prob=.04)
[1] 0.0001368611
```

Good to Know! www.artofstat.com has many very nice online calculators, with wonderful visualizations. One is the Artofstat Binomial Distribution.
www.artofstat.com > Online Web Apps > …. > scroll down …. > Binomial Distribution. At top right, click on the tab *"Find Probability"*. Here is the direct link:

https://istats.shinyapps.io/BinomialDist/

Your Turn

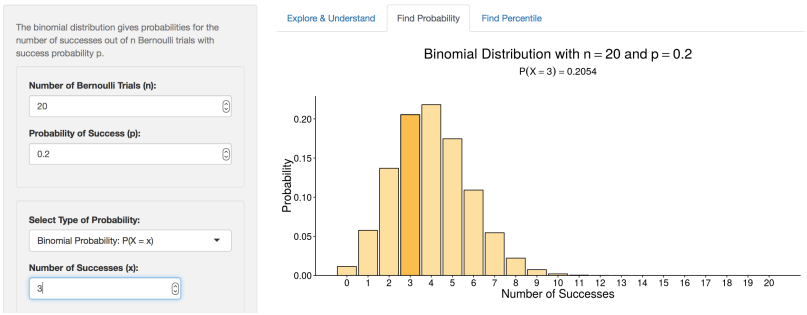
A roulette wheel lands on each of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 with probability = .10. Write down the expression for the calculation of the following.

#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: .2054, representing a 20% chance, approximately.

Online solution - www.artofstat.com > Online WebApps > Binomial Distribution

https://istats.shinyapps.io/BinomialDist/ > (click on tab "Find Probability")	At left Enter number of trials n=20, pr[success]=.20, P(X=x), # successes=3
"Event" is outcome of either "5" or "6" Pr[event] = π = .20 N = 20 X is distributed Binomial(N=20, π=.20) $\Pr[X=3]=\binom{20}{3}[.20]^3[1-.20]^{20-3}$ $=\binom{20}{3}[.20]^3[.80]^{17}$ $=.2054$	The Binomial Distribution 

https://istats.shinyapps.io/BinomialDist/

R Solution - dbinom( ) for exact binomial probabilities

```
# Binomial(n=20, p=.20) Solve for Prob[X=3]
dbinom(x=3, size=20, prob=0.20)
## [1] 0.2053641
```

#2. Solution: .1071, representing an 11% chance, approximately

Online solution - www.artofstat.com > WebApps > Binomial Distribution

https://istats.shinyapps.io/BinomialDist/
> (click on tab "Find Probability")

At left
Enter number of trials n=20, pr[success]=.30, P(X≤x), # successes=3

"Event" is outcome of either "7" or "8" or "9"
Pr[event] = π = .30
N = 20
X is distributed Binomial(N=20, π=.30)

Translation:
"At most 3 times" is the same as "less than or equal to 3 times"

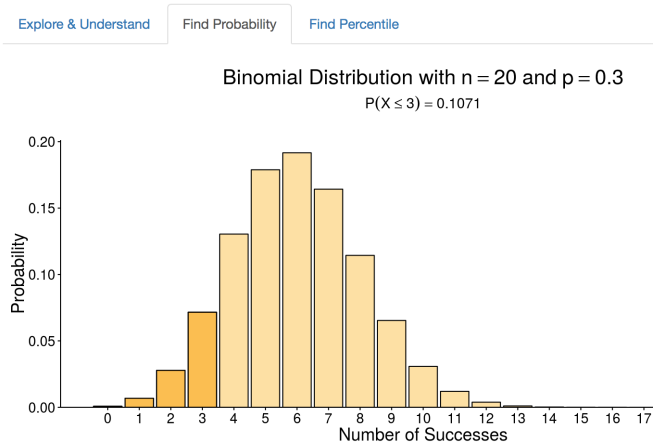
$$\Pr[X \le 3] = \Pr[X=0] + \Pr[X=1] + \Pr[X=2] + \Pr[X=3]$$

$$= \sum_{x=0}^{3} \left\{ \binom{20}{x} \right\} [.30]^x [.70]^{20-x}$$

$$= \binom{20}{0}[.30]^0[.70]^{20} + \binom{20}{1}[.30]^1[.70]^{19}$$

$$+ \binom{20}{2}[.30]^2[.70]^{18} + \binom{20}{3}[.30]^3[.70]^{17}$$

$$= .1071$$



R Solutions **(showing just two here; there are others)**
pbinom( ) for *cumulative* exact binomial probabilities
sum(dbinom( )) for *sum of* exact binomial probabilities

```
# Binomial(n=20, p=.30) Solve for Prob[X <= 3]
pbinom(q=3, size=20, prob=0.30)
## [1] 0.1070868
sum(dbinom(x=0:3, size=20, prob=0.30))
## [1] 0.1070868
```

Review of the Normal Approximation for the Calculation of Binomial Probabilities

Use the Normal Distribution as an Approximation of the Binomial Distribution when:

(1) $n\pi \geq 10$; 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.
Solution = .8146, representing an 81% chance, approximately.

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

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

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

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

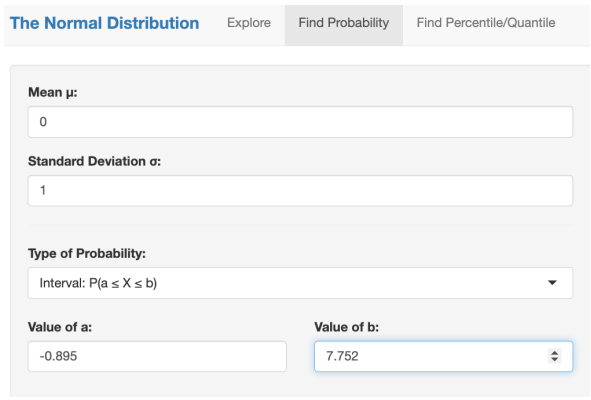
$$= \Pr[\,-.895 \leq \text{Normal}(0,1) \leq 7.752\,]$$

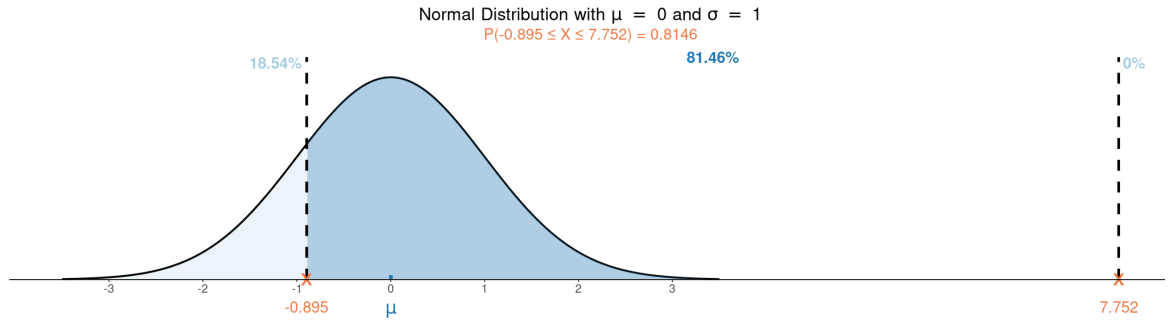
$$\approx \Pr[\,-.895 \leq \text{Normal}(0,1)\,]$$, because 7.752 is in the extreme right tail.

$$= \Pr[\,\text{Normal}(0,1) \leq +.895\,]$$, because of symmetry of the tails of the normal

$$= .8146$$

Online solution - www.artofstat.com > Online WebApps > Normal Distribution
https://istats.shinyapps.io/NormalDist/





Normal Probability (Interval):

μ	σ	a	b	P(a ≤ X ≤ b)
0	1	-0.895	7.752	0.814606

https://istats.shinyapps.io/NormalDist/

R Solution
pnorm( ) - pnorm( ) for area under the curve probabilities

```
> # Pr [ -.985  <=  Normal(0,1)  <= +7.752 ] is the area under the curve between X=-.985 and X=+7.752
> pnorm(q=7.752, mean=0, sd=1) - pnorm(q=-0.895, mean=0, sd=1)
[1] 0.8146065
```

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.

$$\mu_{normal} = \mu_{binomial} = n\pi$$

$$\sigma^2_{normal} = \sigma^2_{binomial} = n\,\pi\,(1-\pi)$$

Desired Binomial Probability Calculation	Normal Approximation with Correction
Pr [X=k]	Pr [(k-1/2) ≤ X ≤ (k+1/2)]
Pr [X > k]	Pr [X > (k-1/2)]
Pr [X < k]	Pr [X < (k+1/2)]

4. Poisson Distribution

When is the Poisson Probability Distribution Model Used? -
The Poisson distribution model is used to investigate rare events and are questions of the form:

- What are the chances of 1 tornado strike hitting Amherst MA in 2024?
- What are the chances of the CDC recording 2 cases of Ebola in the US in 2024?

What is μ in the Poisson Distribution?

μ = expected number of events over a specified frame of observation. The frame might be

* time - a specified total time of surveillance (e.g., 1 year of surveillance)
* Geographic area – a specified area on a map (e.g., continental US) etc.

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 model is a good choice for the modeling of count data that are *rare*.

Tip! In epidemiology, the Poisson distribution model 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 a *"static"* analysis of "x" events among N persons as a proportion.

• Instead, it is an analysis of *"force of events"*, "x" over person time frame, as an incidence rate.

Interesting! The Poisson Distribution is 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 μ,

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

Expected value of X is E[X] = μ
Variance of X is Var[X] = σ2 = μ That's right – mean and variance are the same.

Good to know: The poisson mean μ is the parameter representing the product of the (incidence rate) x (period of observation)

Interpretation of μ – "This is the expected number of events we expect to get over this particular "frame" under this particular Poisson model assumption"

R Users – R refers to the mean μ as "lambda"

The Poisson distribution is an appropriate model for describing the frequency of occurrence of a rare event in a very large (unknowable) number of trials.

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

Solution: .0892, representing a 9% chance approximately.

Step 1: Solve for the Poisson mean parameter μ for the period of observation of interest.

μ = (incidence rate) x (period of observation)
= (2 per 1000 pack years) x (3000 pack years)
= (.002)*(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

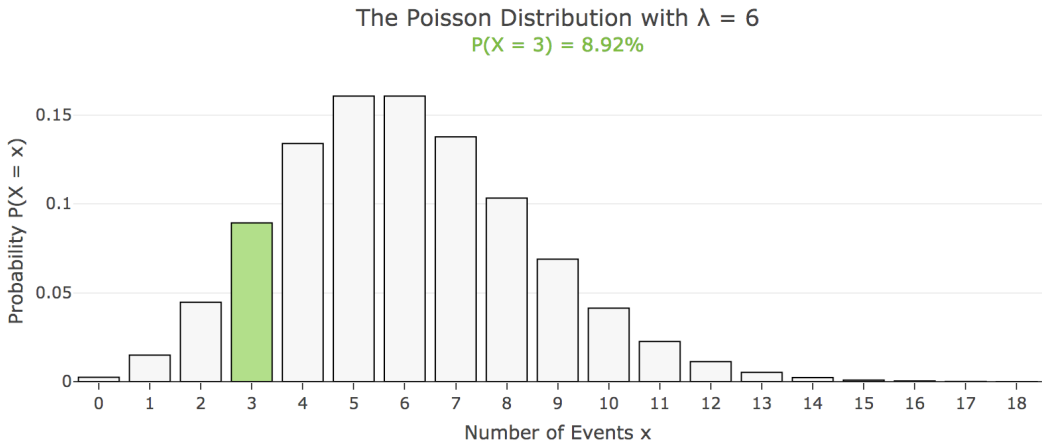
"Calculate the probability of exactly 3 cases of lung cancer" is translated as follows
Here, X= # cases in 3000 pack years and is modeled as distributed Poisson(μ=6),
Probability [X=3] = ???
If we expect 6 under our Poisson model, what are our chances of seeing 3 cases over 3000 pack years?

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.

$$\text{Probability [X=3]} = \frac{\mu^x \exp(-\mu)}{x!} = \frac{6^3 \exp(-6)}{3!} = \frac{6^3 e^{-6}}{3!} = .0892$$

Online solution - www.artofstat.com > Online WebApps > Poisson Distribution
Note: Our notation of the Poisson mean mu is called lambda in this online calculator

https://istats.shinyapps.io/PoissonDist/ > (click on tab "Find Probability")	At left Enter lambda = 6, type of probability: Pr[X=x] and x=3



Poisson Probability

x	P(X = x)
3	0.0892

https://istats.shinyapps.io/PoissonDist/

R Solution - dpois( ) for exact Poisson probabilities
Reminder: Our notation of the Poisson mean mu is called lambda in R

```
#  Poisson(mu=6) Solve for Prob[X=3]

dpois(x=3, lambda=6)          # lambda is the Poisson mean

## [1] 0.08923508
```

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. Note – I did probability calculations by hand here.

	(a) Binomial	**(b) Poisson**
	n = # trials π = Pr [event]	What happens as $n \to \infty$ and $\pi \to 0$?
Expected # events	$n\pi$	μ
Pr [X=x events]	$\binom{n}{x}\pi^x(1-\pi)^{n-x}$	$\frac{\mu^x \exp(-\mu)}{x!}$

where exp = numerical constant = e = 2.718

Solution using the Binomial:

trials = n = 3000

π = .002

Pr[3 cases cancer]

$= \binom{3000}{3}[.002]^3[.998]^{2997}$

= .0891

Solution using the Poisson:

Length of interval, T = 3000

Rate per unit length, λ = .002 per 1 pack year

$\mu = \lambda T$ = (.002/pack year)(3000 pack years)=6

Pr[3 cases cancer]

$= \frac{6^3 \exp[-6]}{3!}$

= .0892

Binomial	**Poisson**
`> dbinom(x=3, size=3000, prob=.002)` `[1] 0.08914568`	`dpois(x=3, lambda=6)` `[1] 0.08923508`

Poisson lambda = (Binomial # Trials) x (Binomial Pr[Event]) = (3000) x (.002) = 6

Cross-Sectional studies in epidemiology are sometimes modeled as sampling from 4 independent Poisson probability distributions

	Disease	Healthy	
Exposed	a	b	a+b
Not exposed	c	d	c+d
	a+c	b+d	a+b+c+d

a = # persons who are both exposed and with disease
b = # persons who are both exposed and healthy
c = # persons without exposure with disease
d = # persons without exposure and healthy

- The counts a, b, c, and d are each separate, independent, random variables. In particular,
- a, b, c, and d are 4 independent Poisson random variables.
- The Poisson means (these are the expected counts) of a, b, c and d are μ_{11}, μ_{12}, μ_{21}, and μ_{22}, respectively.

Likelihood[2x2 table]

= L[a,b,c,d]

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

If there is an association between exposure and disease, what does it look like as regards the Poisson means?
When there is an association,

1) Pr[disease given exposure] ≠ Pr [disease given no exposure]

$$\frac{\mu_{11}}{\mu_{11}+\mu_{12}} \neq \frac{\mu_{21}}{\mu_{21}+\mu_{22}}$$

2) Pr[exposure given disease] ≠ Pr [exposure given no disease]

$$\frac{\mu_{11}}{\mu_{11}+\mu_{21}} \neq \frac{\mu_{12}}{\mu_{12}+\mu_{22}}$$

Two independent groups cohort studies in epidemiology are sometimes modeled as sampling from 2 independent Binomial probability distributions.

	Disease	Healthy	
Exposed	a	b	a+b fixed by design
Not exposed	c	d	c+d fixed by design
	a+c	b+d	a+b+c+d

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 separate independent random variables.
- The counts b and d do not vary because they are obtained by subtraction from the fixed row totals
- a and c are 2 independent Binomial random variables that model the event of disease occurrence
- We might denote the Binomial probability of events parameters (for a and c) π_1 and π_2, respectively

Likelihood[2x2 table]

= L[a,c given (a+b) and (c+d) row totals are fixed]

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

If there is an association between exposure and disease, what does it look like as regards the Binomial event probabilities?
When there is an association,

Pr[disease among exposed] ≠ Pr [disease among non-exposed]

$$\pi_1 \neq \pi_2$$

Two independent groups case-control studies in epidemiology are also sometimes modeled as sampling from 2 independent Binomial probability distributions.

	Disease		Healthy	
Exposed	a	↓	b	a+b
Not exposed	c		d	c+d
	a+c fixed		b+d fixed	a+b+c+d

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 separate independent random variables.
- The counts c and d do not vary because they are obtained by subtraction from the fixed column totals
- a and b are 2 independent Binomial random variables that model the event of history of exposure.
- We might denote the Binomial probability of events parameters (for a and b) θ_1 and θ_2, respectively.

Likelihood[2x2 table]

= L[a,b give (a+c) and (b+d) are fixed]

$$= \left[\binom{a+c}{a}\theta_1^a(1-\theta_1)^c\right]\left[\binom{b+d}{b}\theta_2^b(1-\theta_2)^d\right]$$

If there is an association between exposure and disease, what does it look like as regards the Binomial event probabilities?
When there is an association,

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.

	Spontaneous Abortion	Healthy	
Video Display Terminal	2	21	23
Not	2	234	236
	4	255	259

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] 100% = [0.0154]100% or 1.5%

(2) Assumption of the null hypothesis model says that the overall 1.5% rate applies to both: (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?

Statistical hypothesis testing entails applying the null hypothesis model to the data and examining where it takes us. Under the null hypothesis of "no association", what were the chances that 2 of the 4 events of spontaneous abortions occurred among the exposed? Solution: .038, representing a 4% chance approximately.

"Hand calculation" of the solution. Here, I make 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 the null hypothesis model of "no association", overall:
  # ways to choose 4 abortions from 259 pregnancies is $\binom{259}{4}$ = 183,181,376

- Under the null hypothesis model of "no association", among the 23 "exposed":
- # ways to choose 2 abortions from 23 pregnancies is $\binom{23}{2}$ = 253

- Under the null hypothesis model of "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

$$\frac{\binom{23}{2}\binom{236}{2}}{\binom{259}{4}} = \frac{(253)(27{,}730)}{(183{,}181{,}376)} = .038$$

http://www.mathsisfun.com/combinatorics/combinations-permutations-calculator.html

Solution using R. The function `dhyper( )` is used to obtain exact central hypergeometric distribution probabilities. Take care to use the notation that R uses.

Notation for R Users (n=column 2 total)

	CASES	Non-cases	
EXPOSED = yes	x		k
Not			
	m	n	

R Solution - dhyper( ) for exact central hypergeometric probabilities

```
# Central Hypergeometric(x=#case and exposed, m=#cases, n=#NON-cases, k=#exposed)

# dhyper(x,m,n,k,log=FALSE)
dhyper(2,4,255,23,log=FALSE)

## [1] 0.03829914
```

Refresher – In a 5-card hand of poker, what are the chances of obtaining a hand with exactly 2 queens?
Solution: .0399, representing a 4% chance, approximately.

	Queens	Non-Queens	
Your poker 5 card hand	x = 2		k = 5
Others			
	m = 4	n = 48	

- 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 $\binom{48}{3} = 17{,}296$
- → # 5 card hand with 2 queens and 3 non-queens is $\binom{4}{2}\binom{48}{3} = (6)(17{,}296) = 103{,}776$
- Pr[2 queens in five card hand] $= \dfrac{(\#\text{ of 5 card hands with 2 queens \& 3 non-queens})}{(\#\text{ 5 card hands total})}$

$$= \frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} = \frac{(6)(17{,}296)}{2{,}598{,}960} = .03993$$

http://www.mathsisfun.com/combinatorics/combinations-permutations-calculator.html

R Solution - dhyper( ) for exact central hypergeometric probabilities

```
# Central Hypergeometric (x=#queens in your hand, m=#queens in deck, n=#NON-queens in deck, k=#cards in your hand)
#dhyper(x,m,n,k,log=FALSE)
dhyper(2,4,48,5, log=FALSE)

## [1] 0.03992982
```

The 2x2 Table in Epidemiology - Test of Null: "No Association".

The Null Hypothesis "No Association" model of a 2x2 tables of counts is the Central Hypergeometric.

Often, we are interested in testing the null hypothesis of no association of exposure and disease and the setting is a 2 x 2 table of exposure-disease count data. By convention, the analysis focuses on the the count "a", the number of cases among the exposed.

Question: Is the observed count "a" statistically significantly different from what what is expected under the null hypothesis of no association?

	Case	Control	
Exposed	a	b	(a+b) fixed
Not exposed	c	d	(c+d) fixed
	(a+c) fixed	(b+d) fixed	n = a + b + c + d

Likelihood of 2x2 Table: Cohort Study
Null True: No Association

- COHORT STUDY: 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:
  # 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":
- # 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":
  # 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

$$\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

Nice Result (and for good reason) – We get the same test of no association answer regardless of modeling the 2x2 table of counts as a cohort (prospective) versus case-control (retrospective) study.

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.

Cohort Design	Case-Control Design
Under the null (no association) hypothesis model:	**Under the null (no association) hypothesis model:**
Pr["a" cases among "(a+b)" exposed| fixed marginals] $= \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$	Pr["a" exposed among "(a+c)" cases| fixed marginals] $= \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{a+b+c+d}{a+b}}$

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:

	Disease	Healthy	
Exposed	2	21	23
Not exposed	2	234	236
	4	255	259

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 <u>both</u> exposure and disease
significantly larger than what might have been expected if
there were NO association between exposure and disease?

Recall from page 30 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.

The p-value is obtained by assuming the null hypothesis is true, using this to obtain a null hypothesis model for the chances of our observe data, and then seeing where it takes us. Is our actual data (which is non-negotiable), or any data configuration more unfavorable to the null, likely or unlikely under the null hypothesis model? In particular, what are the null hypothesis model chances (likelihood) of the observed configuration of counts "a", "b", "c", and "d" (or any other configuration that is more extreme) if the row and column totals are held fixed?

Background to the The Fisher Exact Test solution for the p-value. The row and column totals of the 2x2 table are treated as fixed. Because of this, only one cell count can vary; the other cell counts are then determined by subtraction. The resulting model for the 2x2 table of counts with fixed row and column totals is a **hypergeometric** distribution model.

The Fisher Exact Test Null Hypothesis Model. If we now add to this scenario the null hypothesis assumption of "no association", then the model for the 2x2 table of counts with fixed row and column totals is a **central hypergeometric** distribution model. Putting this together:

	Disease	Healthy	
Exposed	a	b	a+b
Not exposed	c	d	c+d
	a+c	b+d	a+b+c+d

Conditional Probability Distribution of "a" in the 2x2 Table
(Row totals fixed and column totals fixed)
Two Scenarios

Scenario 1: Null Hypothesis of "No Association (Odds Ratio, OR = 1)" is Assumed
Tip - We use this model for the calculation of p-values)
When the null is true, the correct model is the model that says "No Association (Odds Ratio, OR = 1)"

Null Hypothesis Model = Central hypergeometric distribution

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

Scenario 2: Alternative Hypothesis of "Any Association (Odds Ratio, OR ≠ 1)" is Assumed
Tip - We use this for sample size and power calculations – not discussed here
When the null hypothesis is NOT true, the correct model is the model that says "Association (Odds Ratio, OR ≠ 1)"

Alternative Hypothesis Model = NON - Central hypergeometric distribution.

Probability [# with exposure and with disease = a]

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

Example - *continued:*
What is the statistical significance (p-value) of 2 abortions under the null hypothesis of "no association"?

Answer: .0410, representing a 4% chance approximately

Solution:
p-value = Probability [2 abortions among the exposed or more extreme relative to the null | null model is true]

= Probability [2 abortions among the exposed | null true]
+ Probability [3 abortions among the exposed | null true]
+ Probability [4 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 if we hold constant 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:
http://easycalculation.com/statistics/odds-ratio.php

a=0

	Disease	Healthy		
Exposed	a = 0	23	23	Pr(table)=.6875
Not exposed	4	232	236	OR=0
	4	255	259	

a=1

	Disease	Healthy		
Exposed	a = 1	22	23	Pr(table)=.2715
Not exposed	3	233	236	OR=3.6
	4	255	259	

a=2 *Note – This is our observed* …

	Disease	Healthy		
Exposed	a = 2	21	23	Pr(table)=.0386
Not exposed	2	234	236	OR=11.1
	4	255	259	

a=3

	Disease	Healthy		
Exposed	a = 3	20	23	Pr(table)=.0023
Not exposed	1	235	236	OR=35.36
	4	255	259	

a=4

	Disease	Healthy		
Exposed	a = 4	19	23	Pr(table)=.0001
Not exposed	0	236	236	OR=infinite
	4	255	259	

Thus,
p-value = Probability [2 abortions among the exposed or more extreme under the null hypothesis model]

= Probability [2 abortions among the exposed | null true]
+ Probability [3 abortions among the exposed | null true]
+ Probability [4 abortions among the exposed | null true]

= .0386
+ .0023
+ .0001

= .0410

R Solutions **(showing just two here; there are others)**
sum(dhyper( )) for sum of *central hypergeometric distributions* exact probabilities
fisher.test( )

```
# p-value calculation using Central Hypergeometric Distribution
> sum(dhyper(x=2:4, m=4, n=255, k=23, log=FALSE))
[1] 0.04062914
```

```
> mytable <- as.table(rbind(c(2, 21),c(2,234)))
> dimnames(mytable) <- list(
+     EXPOSURE=c("Exposed","Not Exposed"),
+     DISEASE=c("Disease","Healthy"))
> mytable
             DISEASE
EXPOSURE      Disease Healthy
  Exposed           2      21
  Not Exposed       2     234


> fisher.test(mytable,alternative="greater")

         Fisher's Exact Test for Count Data

data:  mytable
p-value = 0.04063
alternative hypothesis: true odds ratio is greater than 1
```

Interpretation of Fisher's Exact Test calculations.

The assumption of the null hypothesis of "no association", upon application 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.

Illustration

	Disease	Healthy	
Exposed	a=2	b=21	23
Not exposed	c=2	d=234	236
	4	255	259

R Solution - A Closer Look at the R code.
fisher.test( ) for a 2x2 table (Null: "No Association")

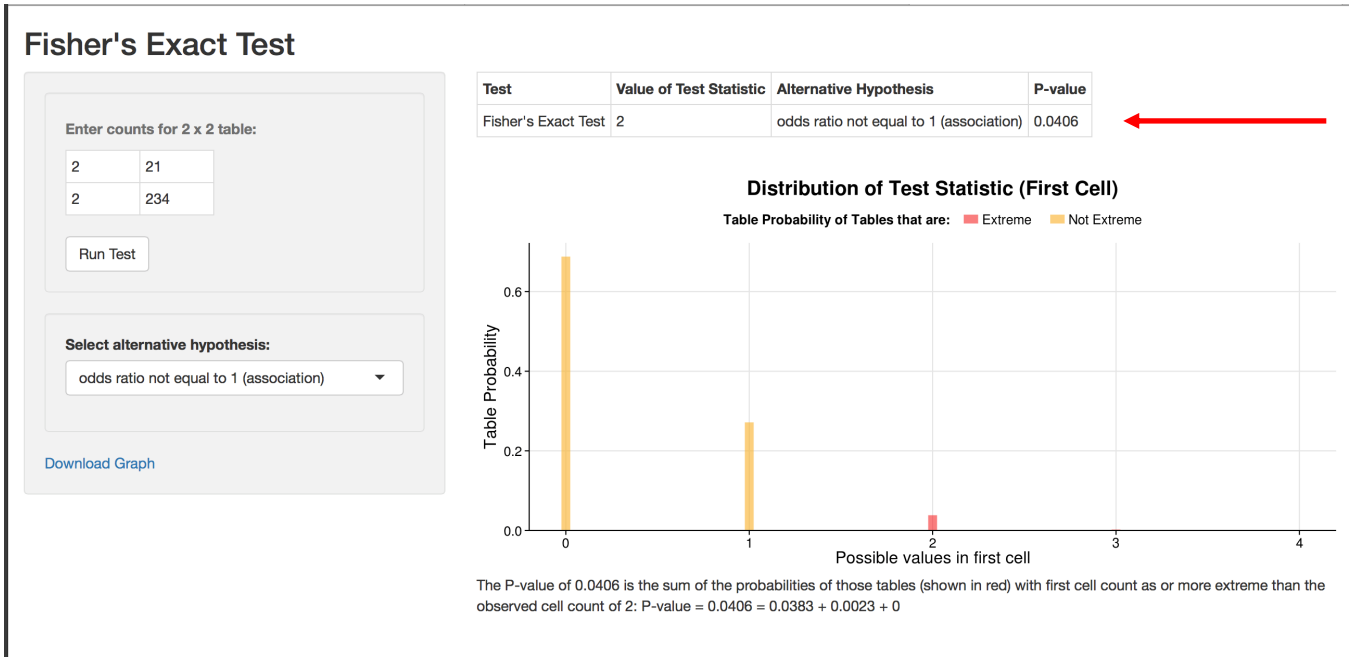
```
# Fisher Exact Test

# STEP 1:  Create 2x2 table using rbind( c( ), c( ), etc ) to bind rows
# mytable <- as.table(rbind(c(a=2,b=21), c(c=2,d=234)))
> mytable <- as.table(rbind(c(2,21), c(2,234)))
> dimnames(mytable) <- list (
      EXPOSURE = c("Exposed", "Not Exposed"),
      DISEASE= c("Disease", "Healthy"))

fisher.test(table2x2, alternative="greater")

##
##  Fisher's Exact Test for Count Data
##
## data:  table2x2
## p-value = 0.04063                          This is of marginal statistical significance (p-value = .04)
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.130931      Inf
## sample estimates:
## odds ratio
##   10.91582
```

Online solution - www.artofstat.com > Online WebApps > Fisher's Exact Test
https://istats.shinyapps.io/FisherExact/



https://istats.shinyapps.io/FisherExact/

7. Discrete Distributions - Themes

	With Replacement	Without Replacement
Framework	A proportion π of the outcomes are events	A fixed population of size =N contains a subset of size =M that are event
Sampling	Sample of size=n with replacement	Sample of size=n without replacement
Outcome	# events = x	# events = x
Likelihood of outcome	$\binom{n}{x}\pi^x(1-\pi)^{n-x}$	$\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$
Example	Test of equality of proportion	Test of independence

Appendix A
Discrete Distribution Calculators

1. Binomial Distribution - *ArtofStat*

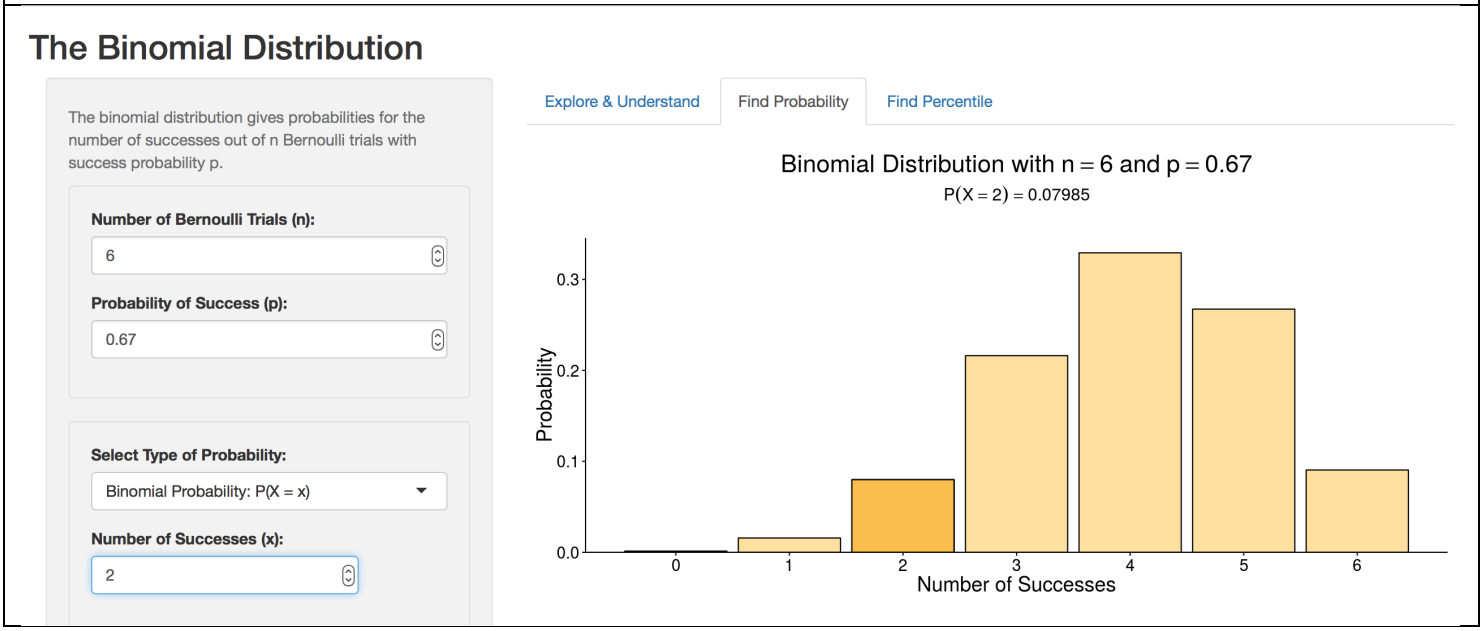
Online using www.artofstat.com > Online Web Apps > scroll down to > BINOMIAL DISTRIBUTION

https://istats.shinyapps.io/BinomialDist/

At top, click on "Find Probability"

From the drop-down menus, select n=# trials, p = Pr[event of success] and type of probability

Example: Fine Pr[X=2] for X ~ Binomial(6, .67) = .07985



1. Binomial Distribution – R

Calculation of Probabilities

	Example: X ~ Binomial (n = 20, p = .20)

Density
Pr [X = x]
`dbinom(x=x, size=, prob=)`
	`> # Pr [ X = 3 ]`
	`> dbinom(x=3,size=20,prob=.20)`
	`[1] 0.2053641`
	`> dbinom(3,20,.20)`
	`[1] 0.2053641`
	`> cat("Pr[Binom(n=20,p=.20) = 3] = ", dbinom(x=3,size=20,prob=.20))`
	`Pr[Binom(n=20,p=.20) = 3] = 0.2053641`

Left Tail
Pr [X <= x]
`sum(dbinom(0:x, size=, prob=))`
`pbinom(q=x, size=, prob=)`
	`> # Pr [ X <= 3 ]`
	`> sum(dbinom(0:3,20,.20))`
	`[1] 0.4114489`
	`> pbinom(q=3,size=20,prob=.20)`
	`[1] 0.4114489`
	`> pbinom(3,20,.20)`
	`[1] 0.4114489`
	`> cat("Pr[Binom(n=20,p=.20) <= 3] = ", pbinom(q=3,size=20,prob=.20))`
	`Pr[Binom(n=20,p=.20) <= 3] = 0.4114489`

Left Tail
Pr [X < x]
`sum(dbinom(0:x-1, size=, prob=))`
`pbinom(q= x-1, size=, prob=)`
	`> # Pr [ X < 3 ] = Pr [ X <= (3-1) ]`
	`> sum(dbinom(0:2,20,.20))`
	`[1] 0.2060847`
	`> pbinom(q=3-1,size=20,prob=.20)`
	`[1] 0.2060847`
	`> pbinom(3-1,20,.20)`
	`[1] 0.2060847`
	`> cat("Pr[Binom(n=20,p=.20) < 3] = ", pbinom(q=3-1,size=20,prob=.20))`
	`Pr[Binom(n=20,p=.20) < 3] = 0.2060847`

Right Tail
Pr [X >= x]
`sum(dbinom(x:n size=, prob=))`
`1 - pbinom(q=x-1, size=, prob=)`
	`> sum(dbinom(3:20,20,.20))`
	`[1] 0.7939153`
	`> 1 - pbinom(q=3-1,size=20,prob=.20)`
	`[1] 0.7939153`
	`> 1 - pbinom(3-1,20,.20)`
	`[1] 0.7939153`
	`> cat("Pr[Binom(n=20,p=.20) >= 3] = ", 1 - pbinom(q=3-1,size=20,prob=.20))`
	`Pr[Binom(n=20,p=.20) >= 3] = 0.7939153`

Right Tail
Pr [X > x]
`sum(dbinom(x+1:n size=, prob=))`
`1 - pbinom(q=x, size=, prob=)`
	`> # Pr [ X > 3 ] = Pr [ X >= 4]`
	`> sum(dbinom(4:20,20,.20))`
	`[1] 0.5885511`
	`> # Pr [ X > 3 ] = 1 - Pr[ X <= 3]`
	`> 1 - pbinom(q=3,size=20,prob=.20)`
	`[1] 0.5885511`
	`> 1 - pbinom(3,20,.20)`
	`[1] 0.5885511`
	`> cat("Pr[Binom(n=20,p=.20) > 3] = ", 1 - pbinom(q=3,size=20,prob=.20))`
	`Pr[Binom(n=20,p=.20) > 3] = 0.5885511`

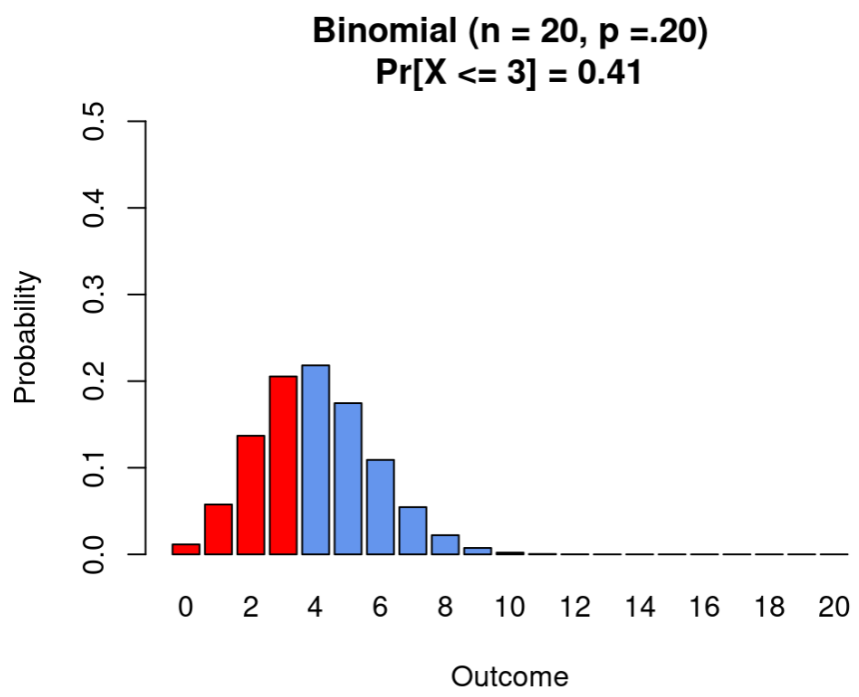
Quantiles `qbinom(p=myquantile, size=, prob=)`	`> qbinom(c(0.025,0.975), size=20,prob=.20)` `[1] 1 8`

1. Binomial Distribution – R

Visualization

```
# Plot Area Under the Curve Corresponding to Pr [ X <= 3 ]

n <- 20                                                  # n = Number of trials
P <- 0.2                                                 # P = Pr [ event ]
likelihoods <- dbinom(x=0:n,size=n, prob=P)              # Get likelihoods Pr[X=0], Pr[X=1], etc
names(likelihoods) <- 0:n                                # Set names for x-axis labeling
cols <- rep("cornflowerblue", n + 1)                     # default color is cornflower blue
cols[1:4] <- "red"                                       # change color to red for X=0,1,2,3
barplot(likelihoods,
          col = cols,
          ylim=c(0,0.5),
          xlab="Outcome",
          ylab="Probability",
          main="Binomial (n = 20, p =.20)\nPr[X <= 3] = 0.41")
```



2. Poisson Distribution - *ArtofStat*

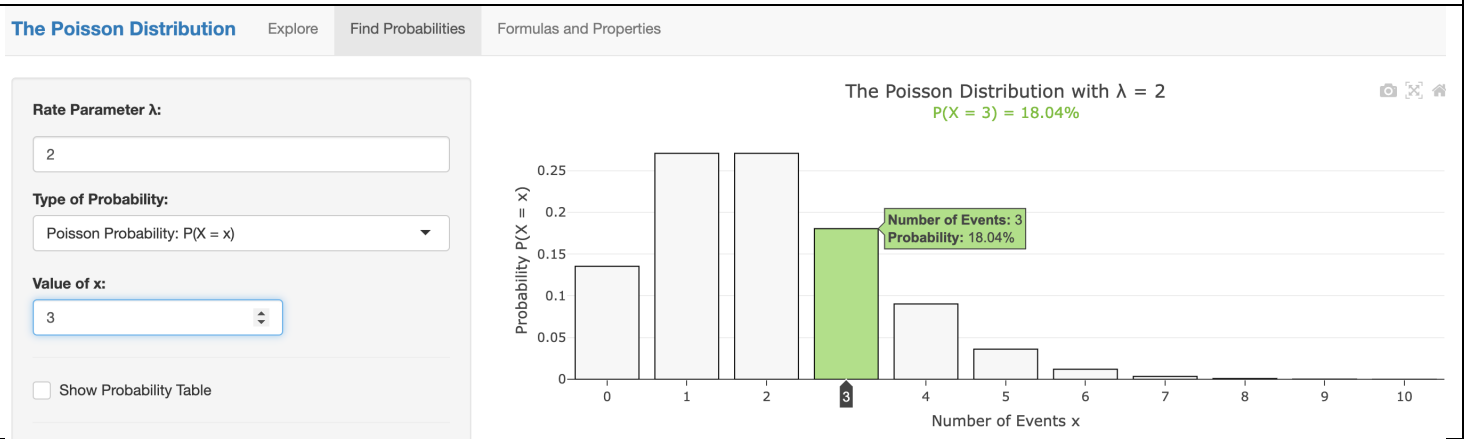
Online using www.artofstat.com > Online Web Apps > scroll down to > POISSON DISTRIBUTION

https://istats.shinyapps.io/PoissonDist/

At top, click on "Find Probabilities"

From the drop-down menus, at box rate parameter λ, enter mean

Example: Fine Pr[X=3] for X ~ Poisson(mean = 2) = .1804



Tip – Try hovering your cursor over the graph! Art of Stat will reward you by telling you what you are looking at!

2. Poisson – R

Calculation of Probabilities

	Example: X ~ Poisson (mean = 2)

Density
Pr [X = x]
`dpois(x=x, lambda=)`
	`> # Pr [ X = 3 ]`
	`> dpois(x=3,lambda=2)`
	`[1] 0.180447`
	`> dpois(3,2)`
	`[1] 0.180447`
	`> cat("Pr[Poisson(mean=2) = 3] = ", dpois(x=3,lambda=2))`
	`Pr[Poisson(mean=2) = 3] = 0.180447`

Left Tail
Pr [X <= x]
`sum(ppois(0:x, size=, lambda=))`
`ppois(q=x, size=, lambda=)`
	`> # Pr [ X <= 3 ]`
	`> sum(dpois(0:3,2))`
	`[1] 0.8571235`
	`> ppois(q=3,lambda=2)`
	`[1] 0.8571235`
	`> ppois(3,2)`
	`[1] 0.8571235`
	`> cat("Pr[Poisson(mean=2) <= 3] = ", ppois(q=3,lambda=2))`
	`Pr[Poisson(mean=2) <= 3] = 0.8571235`

Left Tail
Pr [X < x]
`sum(ppois(0:x-1, size=, lambda=))`
`ppois(q= x-1, lambda=)`
	`> # Pr [ X < 3 ] = Pr [ X <= (3-1) ]`
	`> sum(dpois(0:2,2))`
	`[1] 0.6766764`
	`> ppois(q=3-1,lambda=2)`
	`[1] 0.6766764`
	`> ppois(3-1,2)`
	`[1] 0.6766764`
	`> cat("Pr[Poisson(mean=2) < 3] = ", ppois(q=3-1,lambda=2))`
	`Pr[Poisson(mean=2) < 3] = 0.6766764`

Right Tail
Pr [X >= x]
`1 - ppois(q=x-1, lambda=)`
	`> # Pr [ X >= 3 ] = 1 - Pr[ X <= (3-1)]`
	`> 1 - ppois(q=3-1,lambda=2)`
	`[1] 0.3233236`
	`> 1 - ppois(3-1,2)`
	`[1] 0.3233236`
	`> cat("Pr[Poisson(mean=2) >= 3] = ", 1 - ppois(q=3-1,lambda=2))`
	`Pr[Poisson(mean=2) >= 3] = 0.3233236`

Right Tail
Pr [X > x]
`1 - ppois(q=x, lambda=)`
	`> # Pr [ X > 3 ] = 1 - Pr[ X <= 3]`
	`> 1 - ppois(q=3,lambda=2)`
	`[1] 0.1428765`
	`> 1 - ppois(3,2)`
	`[1] 0.1428765`
	`> cat("Pr[Poisson(mean=2) > 3] = ", 1 - ppois(q=3,lambda=2))`
	`Pr[Poisson(mean=2) > 3] = 0.1428765`

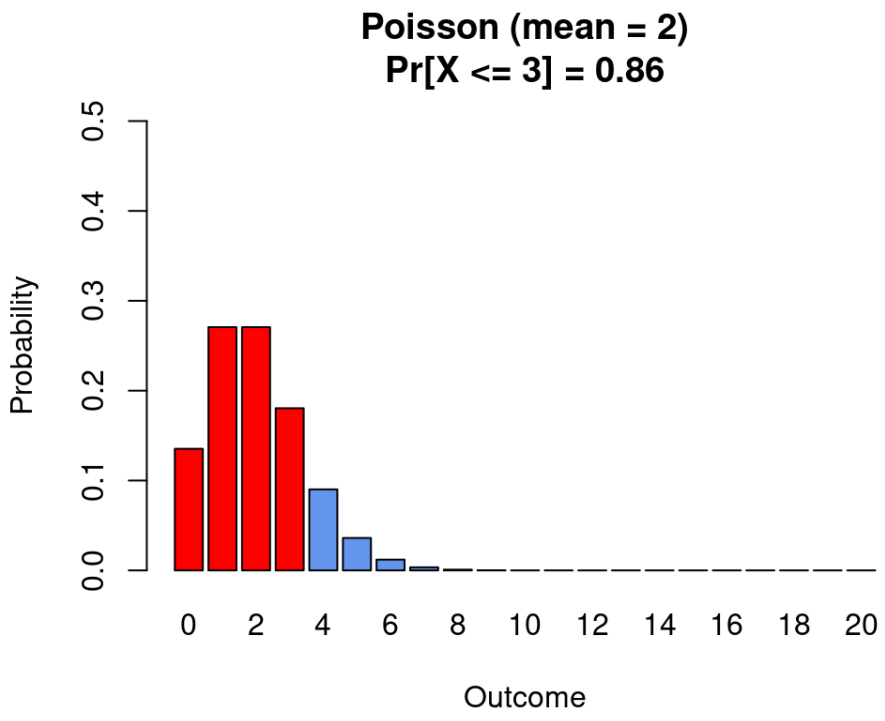
Quantiles
`qpois(p=myquantile, lambda=)`
	`> # 2.5th and 97.5th Quantiles`
	`> qpois(c(0.025,0.975), lambda=2)`
	`[1] 0 5`

2. Poisson – R

Visualization

```
# Plot Area Under the Curve Corresponding to Pr [ X <= 3 ]

> # Pr [ X <= 3 ] Graph
> mymean <- 2                                       # mymean = desired mean of Poisson
> likelihoods <- dpois(x=0:20,lambda=mymean)        # get likelihoods for desired values of x
> names(likelihoods) <- 0:20                        # set names for x-axis labeling
> cols <- rep("cornflowerblue", 21)                 # default color is cornflower blue
> cols[1:4] <- "red"                                # change color to red for X=0,1,2,3
>   barplot(likelihoods,
+           col = cols,
+           ylim=c(0,0.5),
+           xlab="Outcome",
+           ylab="Probability",
+           main="Poisson (mean = 2)\nPr[X <= 3] = 0.86")
```



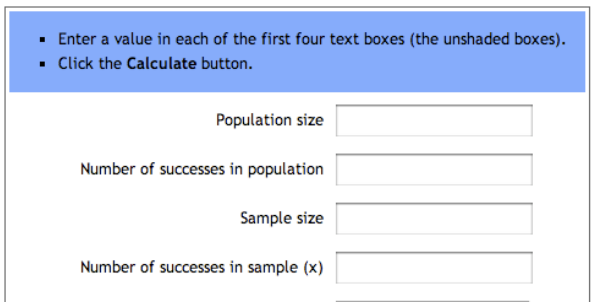
3. (Central) Hypergeometric Distribution - *Stattrek*

Case and exposed # = 2

	CASES	Non-cases	
Exposed = yes	a=2	b=21	23 = Exposed, total
Not exposed	c=2	d=234	236
	4	255	**n=259**
	= Cases, total		= Sample size, total

Online

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



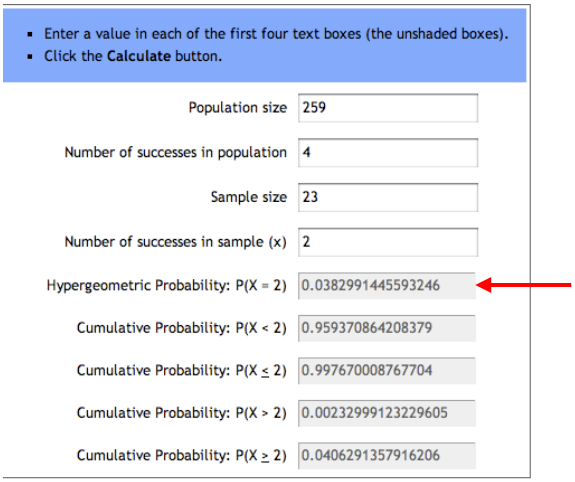
Definition of Boxes at Left. How to enter values:

"Population size" = n = total of 2x2 table
"Number of successes" = # Cases, total = (a+c) = column 1 total
"Sample size" = # Exposed, total = (a+b) = row 1 total
"Number of successes in sample" = # Exposed AND Case = 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] = .038299


Enter a value in each of the first four text boxes (the unshaded boxes).
Click the Calculate button.
Population size 259
Number of successes in population 4
Sample size 23
Number of successes in sample (x) 2
Hypergeometric Probability: P(X = 2) 0.0382991445593246
Cumulative Probability: P(X < 2) 0.959370864208379
Cumulative Probability: P(X ≤ 2) 0.997670008767704
Cumulative Probability: P(X > 2) 0.00232999123229605
Cumulative Probability: P(X ≥ 2) 0.0406291357916206

3. (Central) Hypergeometric Distribution - *R*

Take care! Take care to use the notation that R uses.

Notation for R Users (n=column 2 total)

	CASES	Non-cases	
EXPOSED = yes	x = 2	21	k=23
Not	2	234	236
	m =4	n=255	**259**

```
# Key:
# x = #cases among exposed (this is "a" in epidemiology parlance)
# m = #cases (this is also the column 1 total, or "a + c" in epidemiology parlance)
# n= # NON-cases (this is also the column 2 total, or "b + d" in epidemiology parlance)
# k=#exposed (this is the row 1 total, or "a + b" in epidemiology parlance)

# R command dhyper( )
# dhyper(x=, m=, n=, k= , log=FALSE)
```

```
>   # Calculate Pr [ X = 2 ]:

>   dhyper(x=2,m=4,n=255,k=23,log=FALSE)
[1] 0.03829914

>   dhyper(2,4,255,23,log=FALSE)
[1] 0.03829914

>   cat("Pr[ X = 2] = ",dhyper(x=2,m=4,n=255,k=23,log=FALSE) )
Pr[ X = 2] =  0.03829914
```

Appendix B

Mean (μ) and Variance (σ^2) of a Bernoulli Distribution

Mean of Z = $\mu = \pi$

The mean of Z is represented as E[Z]. recall: "E" stands for "statistical expection value of"

E[Z] = π 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 = $\sigma^2 = (\pi)(1-\pi)$

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

Var[Z] = $\pi(1-\pi)$ because the following is true:

$$\text{Var}[Z] = E[(Z-\pi)^2] = \sum_{\text{All possible z}} (z-\pi)^2\ \text{Probability}[Z=z]$$

$$= [(0-\pi)^2]\Pr[Z=0] + [(1-\pi)^2]\Pr[Z=1]$$

$$= [\pi^2]\ (1-\pi) + [(1-\pi)^2]\ (\pi)$$

$$= \pi\ (1-\pi)\ [\ \pi + (1-\pi)\ ]$$

$$= \pi\ (1-\pi)$$

Appendix C
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.

- $\sum_{i=1}^{n} Z_i = X = \#$ events in n trials

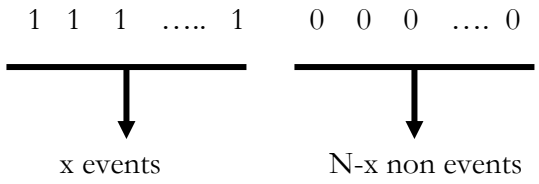
This new random variable X is distributed Binomial.
X represents the net number of successes in a set of independent Bernoulli trials. A simple example is the outcome of several coin tosses (eg – *"how many heads did I get?"*). The word choice "net" is deliberate here, to remind ourselves that we're 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 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.



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

$$= \pi\ \pi\ \pi \ldots \pi\ (1-\pi)\ (1-\pi)\ (1-\pi) \ldots (1-\pi)$$

$$= \pi x \quad (1-\pi)n-x$$

Note! Can you see that the result is the same product $\pi x \quad (1-\pi)n-x$ 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.

$$\text{Number of } \textit{"qualifying"} \text{ ordered sequences} = \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

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])

$$\Pr[X = x] = \Pr\left[\sum_{i=1}^{n} Z_i = x\right] = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

Appendix D
The Poisson Distribution is an Extension of the 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 = total accumulation of person time (e.g. – 1000 pack years)
  n = number of sub-intervals of T (eg – 1000)
  T/n = length of 1 sub-interval of T (e.g.- 1 pack year)
  λ = event rate per unit length of person time

- What is our Binomial distribution probability parameter π?

  $\pi = \lambda (T/n)$ because it is (rate)(length of 1 sub-interval)

- What is our Binomial distribution number of trials?

  n = 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 = Pr[1 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] = $\binom{n}{x}$ π^x $(1-\pi)^{n-x}$ for x=0, …, n

$$= \binom{n}{x}\left[\frac{\lambda T}{n}\right]^x\left[1-\frac{\lambda T}{n}\right]^{n-x} \text{ because } \pi=(\lambda)[T/n].$$

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 λT because λ 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 -

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

As n→infinity, the product of terms in the numerator → (n)(n)(n) … (n) = nx.

Thus, as n → infinity

$$\binom{n}{x} \rightarrow \frac{n^x}{x!}$$

- 2nd 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,

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

- 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→infinity $\left[1-\frac{\mu}{n}\right]^n \rightarrow e^{-\mu}$ where e = constant = 2.718

Thus, as n → infinity

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

- 4th term -

Finally, as n→ the quotient $(\lambda T/n)$ is increasingly like (0/n) so that

$$\left[1-\frac{\lambda T}{n}\right]^{-x} \rightarrow \left[1-\frac{0}{n}\right]^{-x} \rightarrow [1]^{-x} = 1$$

Thus, as n → infinity

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

Now put together the product of the 4 terms and what happens as n→ infinity

$$\Pr[X=x] = \binom{n}{x}\left[\frac{\lambda T}{n}\right]^{x}\left[1-\frac{\lambda T}{n}\right]^{n}\left[1-\frac{\lambda T}{n}\right]^{-x} \xrightarrow{\text{as } n\rightarrow \text{ infinity}}$$

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

$$= \frac{\mu^{x}\, e^{-\mu}}{x!} = \frac{\mu^{x}\, \exp^{-\mu}}{x!}$$

Poisson Distribution

If X is distributed Poisson (μ),

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

Expected value of X is $E[X] = \mu$
Variance of X is $Var[X] = \sigma 2 = \mu$

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