Logistic Regression for Count and Proportion Data

Karen Grace-Martin

A Motivating Example

SubID	Age	Trt_Group	Num_Symptoms	Total	Prop_Symptoms	Complication
1	52	0	20	20	1.00	1
2	67	0	11	20	.55	1
3	59	1	9	20	.45	0
4	68	0	15	20	.75	1
5	67	1	15	18	.83	0
6	69	0	15	20	.75	0
7	72	1	14	20	.70	1
8	71	0	0	20	.00	0
9	65	1	18	20	.90	0
10	78	0	1	20	.05	0
11	72	1	12	20	.60	1
12	68	0	13	20	.65	1
13	70	1	0	20	.00	1
14	52	0	8	17	.47	0

What You'll Learn Today

- Review Linear and Binary Logistic Regression
- The Bernoulli and Binomial Distributions
- Logistic Regression for Binomial Data:
  What does and doesn't work

Brief Review of Linear and Logistic Regressions

The Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{k1} X_{ki} + \varepsilon_i$$

- Y is the response variable
- X_j is the j^{th} predictor variable
- β_0 is the Y-intercept
- β_j is the coefficient of the j^{th} predictor variable
- ε is the residual error
- Y_i and X_i are the values of X and Y for the i^{th} individual

Distributional Assumptions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{k1} X_{ki} + \varepsilon_i$$

$\varepsilon_i \sim iid\ N(0, \sigma^2)$

The variance of the errors (and Y|X) is constant

The errors (and Y|X) are normally distributed

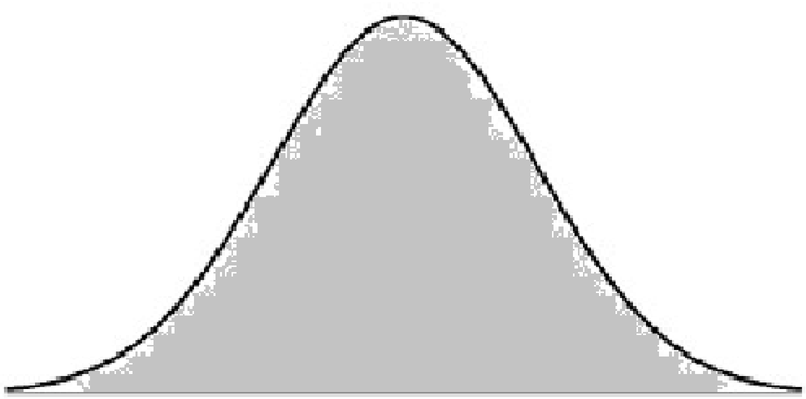
The errors (and Y|X) are independent of each other

The Normal Distribution

Truly Numerical

Truly Continuous

Range is -∞ to ∞

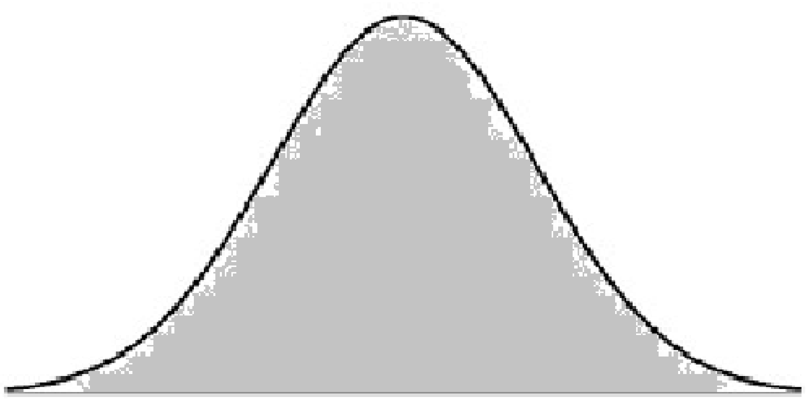


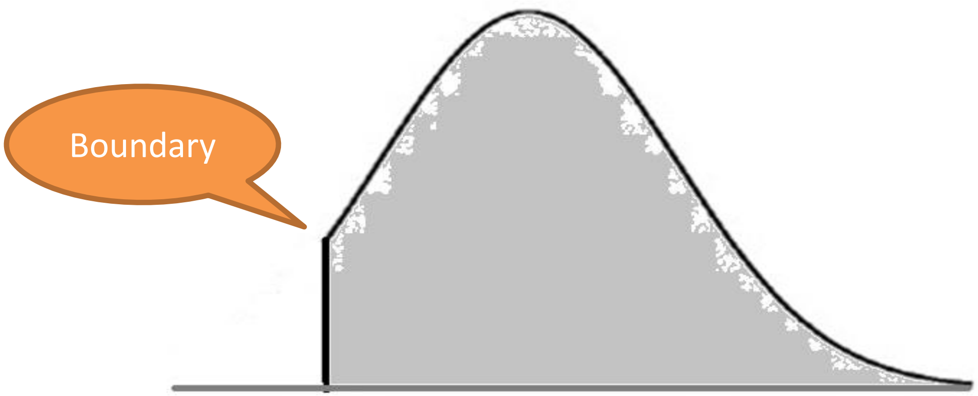
The Normal Distribution

Truly Numerical

Truly Continuous

Range is -∞ to ∞





The Distribution of Y

Y cannot be:

- categorical
- ordinal
- discrete counts
- zero inflated
- censored or truncated, including time to event
- bounded, including a proportion or percentage

Binary Response Variable

Y = 1 if a student passes a class

Y = 0 if a student does not pass a class

Y = 1 if a frog lives

Y = 0 if a frog dies

Y = 1 if a second-language learner makes a grammar mistake

Y = 0 if a second-language learner does not make a grammar mistake

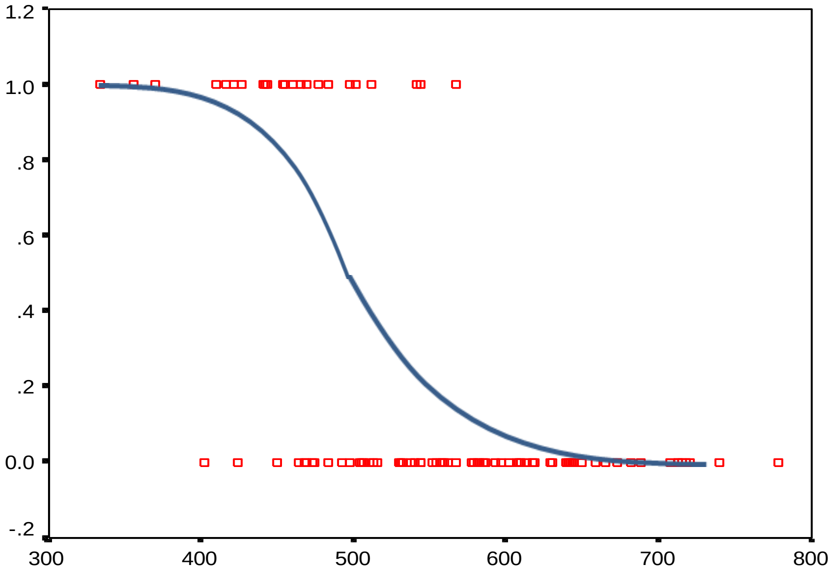
Binary Logistic Regression

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

- X_j is the j^{th} predictor variable
- β_0 is the Y-intercept
- β_j is the coefficient of the j^{th} predictor variable

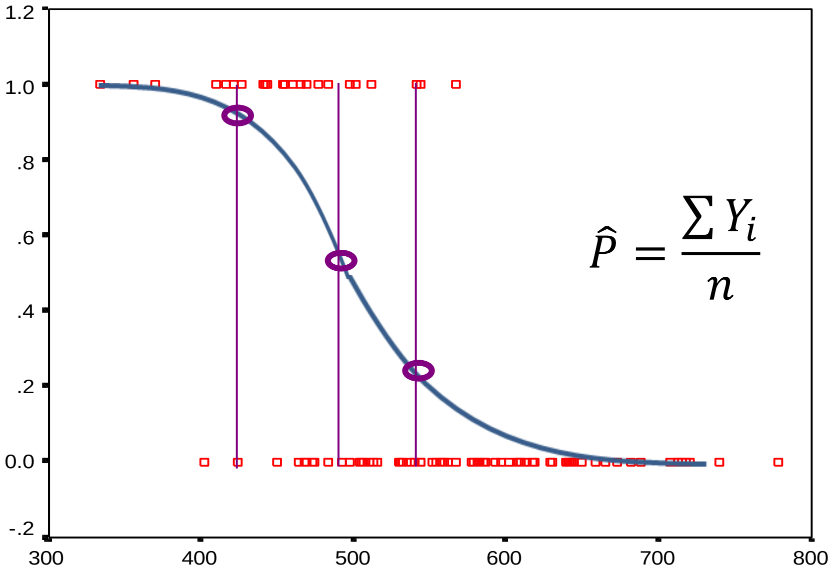
P is probability Y=1 on each trial

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{k1} X_{ki} + \varepsilon_i$$



Doesn't work

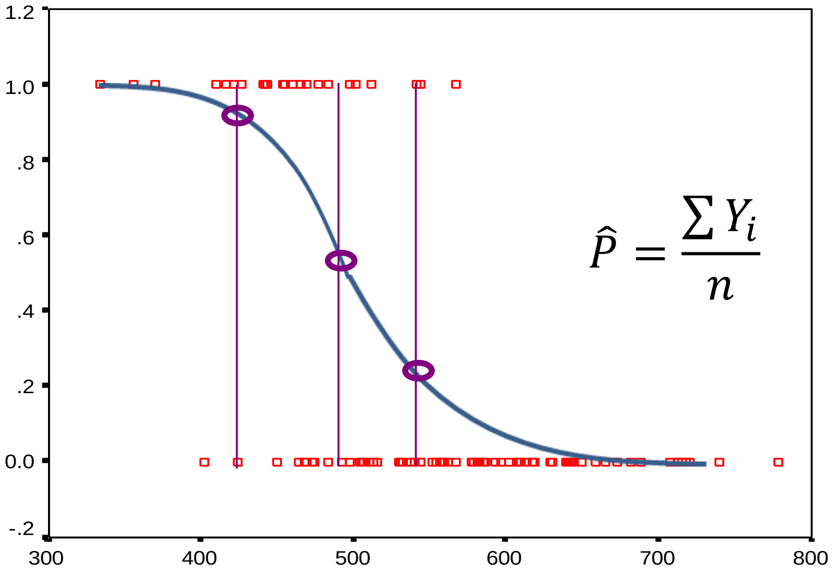
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{k1} X_{ki} + \varepsilon_i$$



$$\hat{P} = \frac{\sum Y_i}{n}$$

Works!

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$



$$\hat{P} = \frac{\sum Y_i}{n}$$

The Bernoulli and Binomial Distributions

Bernoulli Trials

A trial with one of two possible outcomes:

1 = Success (the outcome you're interested in tracking)

0 = Failure

Such that:

$P(1) = p$

$P(0) = 1-p$

Bernoulli Distribution

Let Y be a discrete random variable that represents the outcome of a Bernoulli trial

Y ~ Bernoulli (p)

Probability Mass Function:

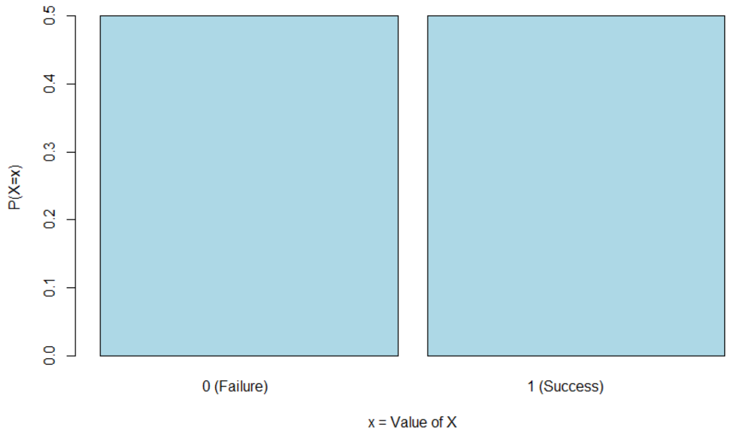
$$P(Y = y) = \begin{cases} 1 - p & for\ y = 0 \\ p & for\ y = 1 \end{cases}$$

Moments:

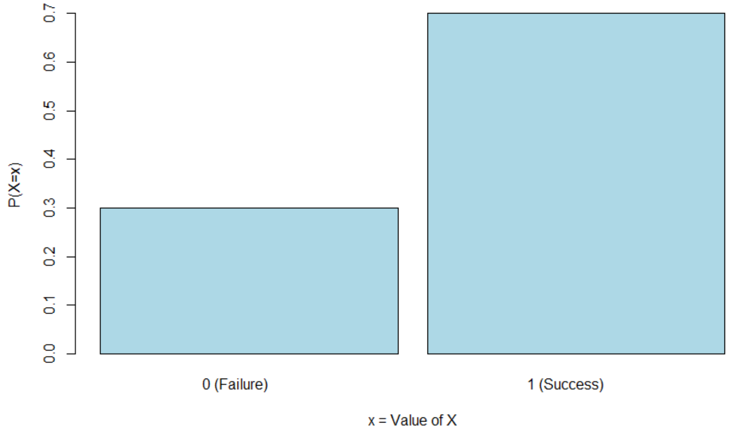
$\mu(Y) = p$

$Var(Y) = p(1-p)$

Bernoulli Distribution



p = .5



p = .7

Sets of Bernoulli Trials

Each trial must be independent

Each trial must have the same probability of success, p

The Binomial Distribution

Let Y be a random variable that represents the number of successes in a set of Bernoulli trials that is repeated n times

N = number of Bernoulli trials

P = probability of success on each Bernoulli trial

Y = number of successes

The Binomial Distribution

Criteria:

1. The number of trials is fixed
2. Each trial is independent
3. The probability of success, p, is the same on each trial

The Binomial Distribution

Examples

Y = The number of classes a student passed each semester out of all classes taken

Y = The number of frogs who survived out of the total number of frogs

Y = The number of sentences on which the second-language learner made a particular grammatical error out of 20 sentences

The Binomial Distribution

If random variables Y_1, Y_2, … Y_k ~ *iid* Bernoulli (p)

Then $\sum_{k=1}^{n}(Y_i)$ ~ Binomial(n,p)

Example:

Let Y_1 measure passing sociology
Let Y_2 measure passing macroeconomics
Let Y_3 measure passing physics
Let Y_4 measure passing Spanish 2

$$P(Y_i = y) = \begin{cases} .2 & for\ y = 0 \\ .8 & for\ y = 1 \end{cases}$$

Then Y = $Y_1 + Y_2 + Y_3 + Y_4$ the number of courses passed this semester, out of four possible classes

The Binomial Distribution

n = number of Bernoulli trials

p = probability of success on each Bernoulli trial

Y = number of successes

Probability Mass Function:

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

Moments:

$\mu(Y) = np$

$Var(Y) = np(1-p)$

The Binomial Distribution

Probability Mass Function:

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

Moments:

$\mu(Y)$ = np

Var(Y) = np(1-p)

Example: What is the probability of passing exactly 3 classes in a semester if the probability of passing each class is .8?

The Binomial Distribution

Probability Mass Function:

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

Moments:

$\mu(Y)$ = np

Var(Y) = np(1-p)

Example: What is the mean number of classes passed out of 4 if the probability of passing each class is .8?

The Binomial Distribution

Probability Mass Function:

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

Moments:

μ(Y) = np

Var(Y) = np(1-p)

Example: What is the probability of a passing each class if the mean number of classes passed out of 4 is 3.6?

Two ways to estimate p, the probability of success on each trial

Bernoulli data set

Each student takes one class

Each class is a Bernoulli trial

Class is the unit of analysis

Y = 1 or 0

Binomial data set

Each student takes 4 classes

Each class is a Bernoulli trial

Student is the unit of analysis

Y is the number of classes passed out of 4

Application back to logistic regression

Bernoulli data set

Each subject has one Bernoulli trial

The Bernoulli trial is the unit of analysis for the outcome variable

Values of Y are 1 or 0

All predictors are measured at the subject level (which is the same as the trial level)

Binomial data set

Each subject has multiple Bernoulli trials

The subject is the unit of analysis for the outcome variable

Values of Y are a count or proportion of successes

All predictors are measured at the subject level (which is NOT the same as the trial level)

Logistic Regression for Binomial Data: What does and doesn't work

Data Set up

SubID	Age	Trt_Group	Num_Symptoms	Total	Prop_Symptoms	Complication
1	52	0	20	20	1.00	1
2	67	0	11	20	.55	1
3	59	1	9	20	.45	0
4	68	0	15	20	.75	1
5	67	1	15	18	.83	0
6	69	0	15	20	.75	0
7	72	1	14	20	.70	1
8	71	0	0	20	.00	0
9	65	1	18	20	.90	0
10	78	0	1	20	.05	0
11	72	1	12	20	.60	1
12	68	0	13	20	.65	1
13	70	1	0	20	.00	1
14	52	0	8	17	.47	0

Writing the model

SubID	Age	Trt_Group	Num_Symptoms	Total	Prop_Symptoms	Complication
1	52	0	20	20	1.00	1
2	67	0	11	20	.55	1
3	59	1	9	20	.45	0
4	68	0	15	20	.75	1
5	67	1	15	18	.83	0

$Y = X_1 \; X_2$

Binary: `Complication = Trt_Group Age`

Binomial: `Num_Symptoms/Total = Trt_Group Age`

Writing the model

SubID	Age	Trt_Group	Num_Symptoms	Total	Prop_Symptoms	Complication
1	52	0	20	20	1.00	1
2	67	0	11	20	.55	1
3	59	1	9	20	.45	0
4	68	0	15	20	.75	1
5	67	1	15	18	.83	0

$$Y = X_1 \; X_2$$

Binary: `Complication = Trt_Group Age`

Binomial: `Num_Symptoms/Total = Trt_Group Age`

Events

Trials

Do you really need GLMM or GEE?

We can use a binomial model only if:

- p is the same on every trial
- trials are independent

$\mu(Y) = np$
$Var(Y) = np(1-p)$

Overdispersion results when the n_i Bernoulli trials per group are

- not identically distributed or
  - (i.e. p varies across trials)
- not independent
  - the outcome of one trial influences the outcomes of other trials
  - The groups of trials are actually clusters with different p

Proportions and Percentages that Can't use Logistic Regression

Continuous proportions

- Space
- Time
- Composition

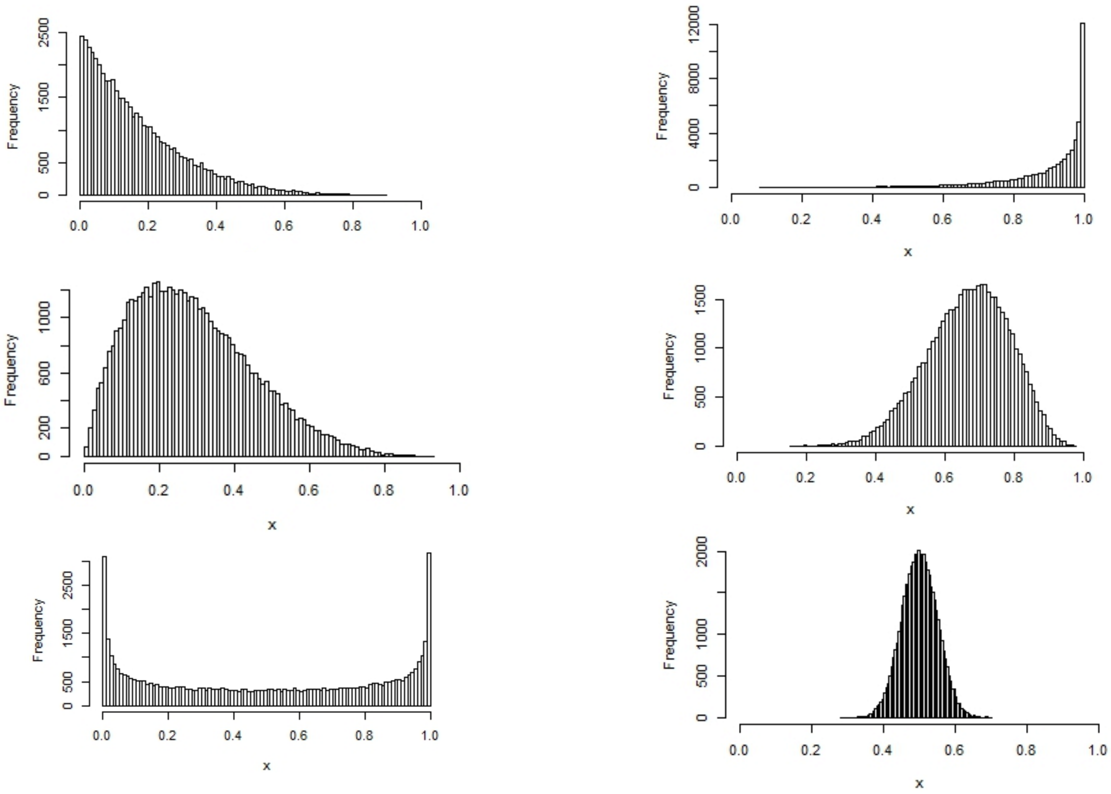
Proportions and Percentages that Can't use Logistic Regression

Continuous proportions

- Space
- Time
- Composition

Beta Distribution

0 < Y < 1

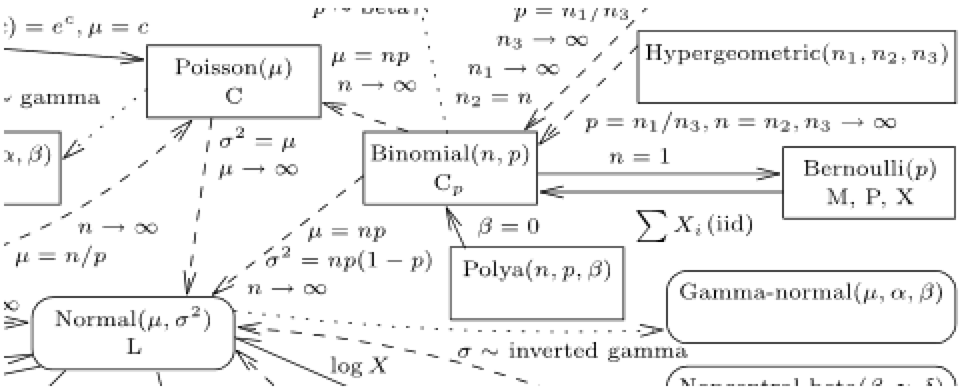


Counts that Can't use Logistic Regression

1. Uncountable or infinite number of trials

- How many frogs were in each pond?

- Number of employees in a state with injury



Counts that Can't use Logistic Regression

1. Uncountable or infinite number of trials

- How many frogs were in each pond?
- Number of employees in a state with injury

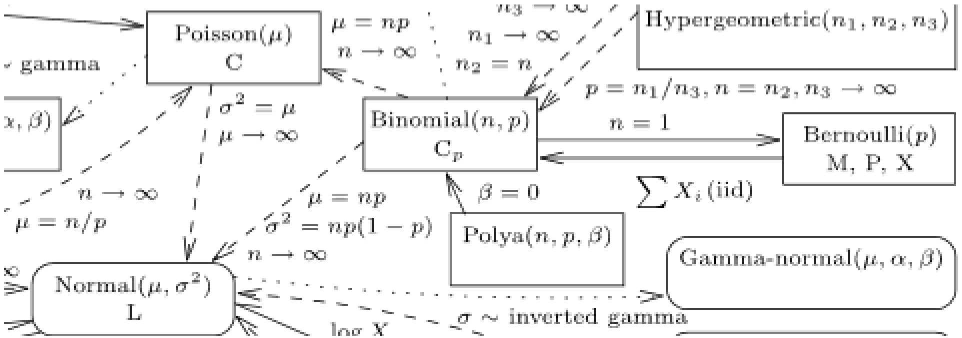
2. No clear trials

- Number of days in the hospital
- Number of grammar mistakes in a document

Counts and Percentages that CAN use a Normal Distribution

1. When a binomial has a large n and a p in the middle (between .2 and .8)
2. When a Poisson variable has a large μ

Make sure assumptions are met



References

J. Tanton. (2017) Curriculum Essay_Poisson Distribution

http://www.jamestanton.com/wp-content/uploads/2012/03/Curriculum-Essay_December-2017_Poisson-Distribution.pdf

L. Leemis & J. McQueston (2008). Univariate Distribution Relationships. The American Statistician.
http://www.math.wm.edu/~leemis/2008amstat.pdf

More on Overdispersion: https://onlinecourses.science.psu.edu/stat504/node/162/

R. Larsen & M. Marx (1986). An Introduction to Mathematical Statistics and Its Applications, 2nd Ed. Prentice Hall.