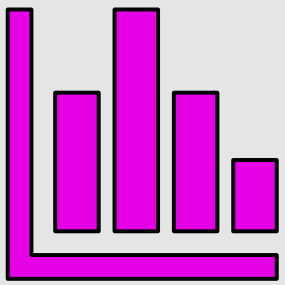
The Wide World of Distributions

BERDC Special Topics Talk 7



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson

Biostatistics, Epidemiology, and Research Design Core

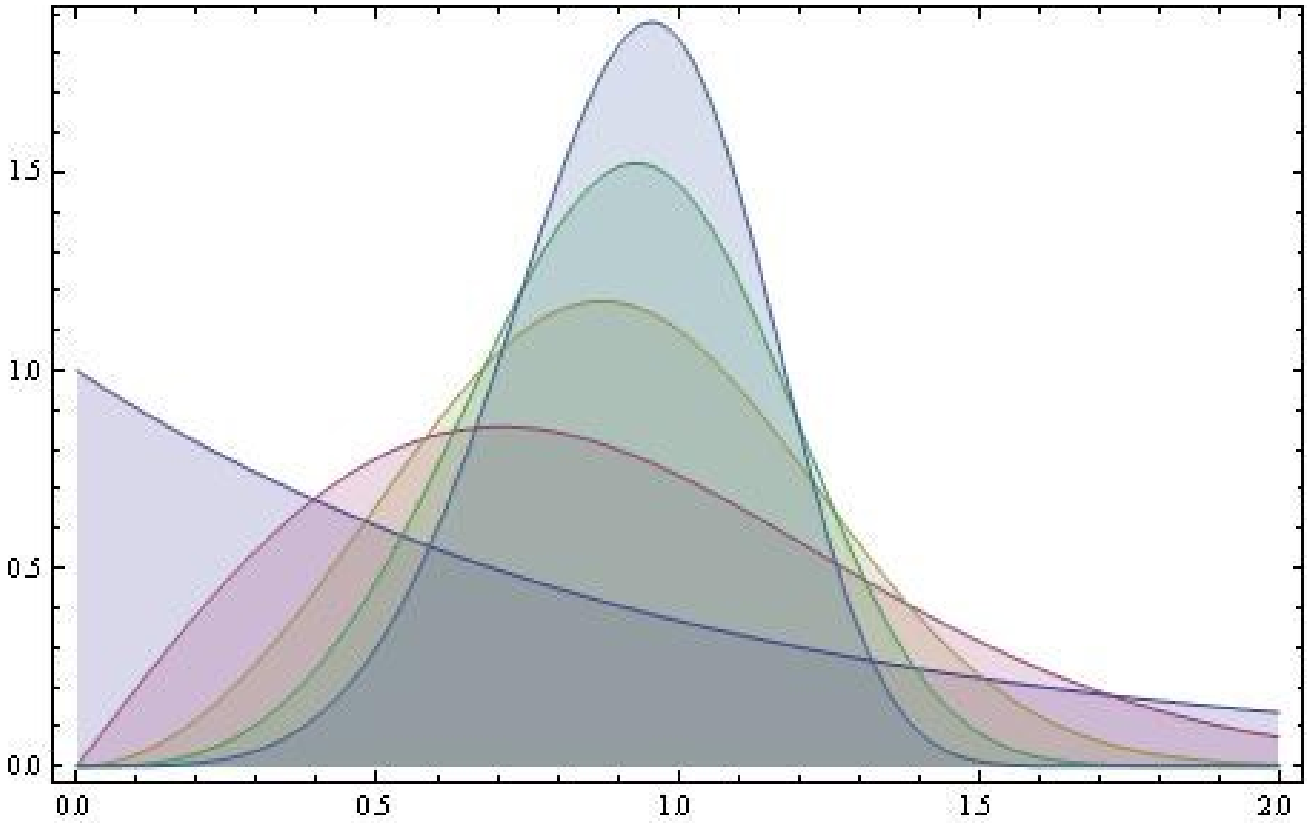
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Opening

UND UNIVERSITY OF NORTH DAKOTA.

Goal: learn about statistical distributions

- Better statistical understanding
- Better statistical modeling
- Better statistical insights



Before Moving On:

Pre-test: https://und.qualtrics.com/jfe/form/SV_1QRRQKFLG086iuW

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

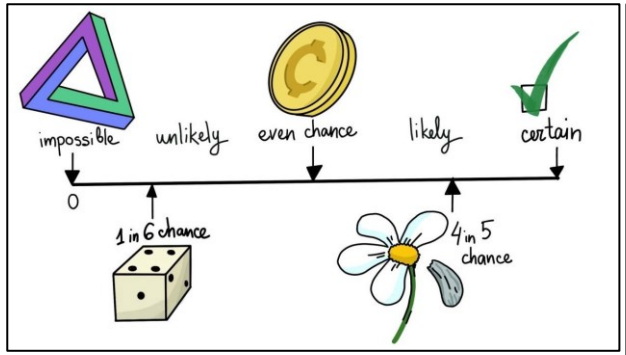
Definition

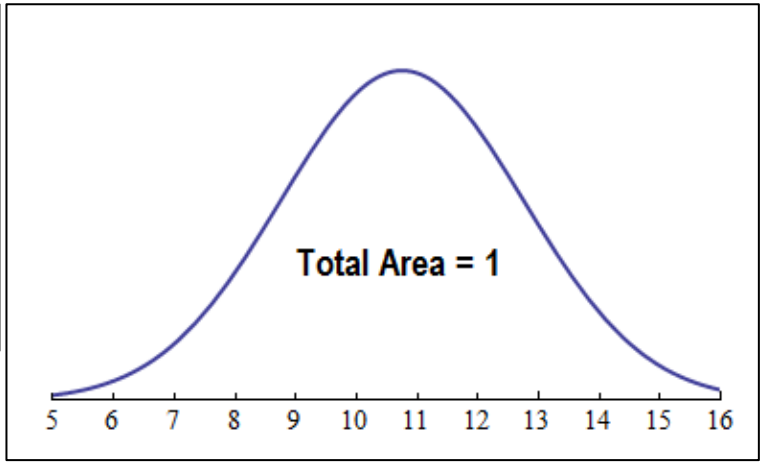
UNIVERSITY OF NORTH DAKOTA

"Function that shows the possible values for a variable and how often they occur"

Properties

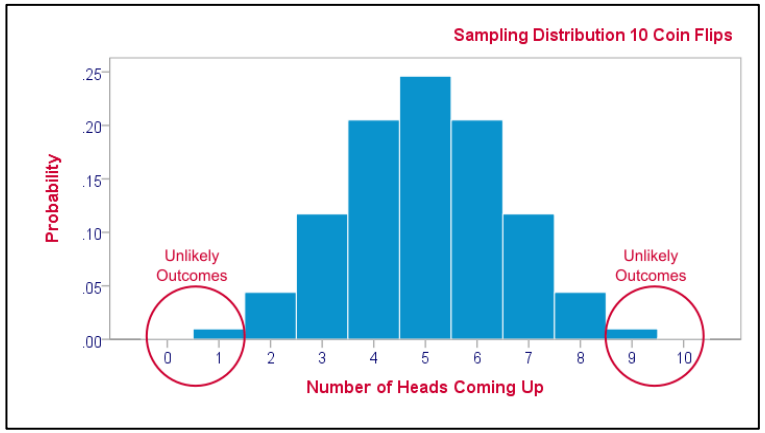
- probability that x can take a specific value is p(x)
- p(x) is non-negative for all real x
- sum of p(x) over all x equals 1





From the variable's perspective

- variables are sampled from random distributions that occur naturally
- each random distribution tells a different story about the processes that produce that variable
- each distribution has a mean, variance, and probability density function



"Mathematical construct to describe how variables are generated"

"Mathematical description of how data conceivably can be produced"

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Variable origins

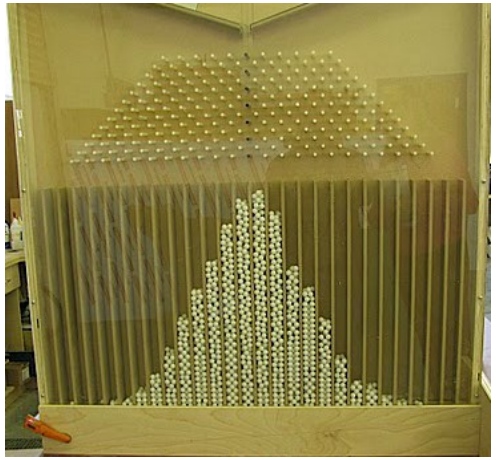
UND UNIVERSITY OF NORTH DAKOTA.

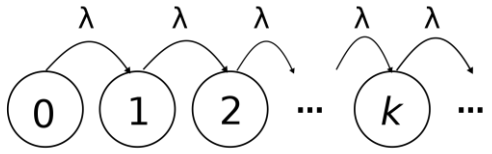
Random generation processes

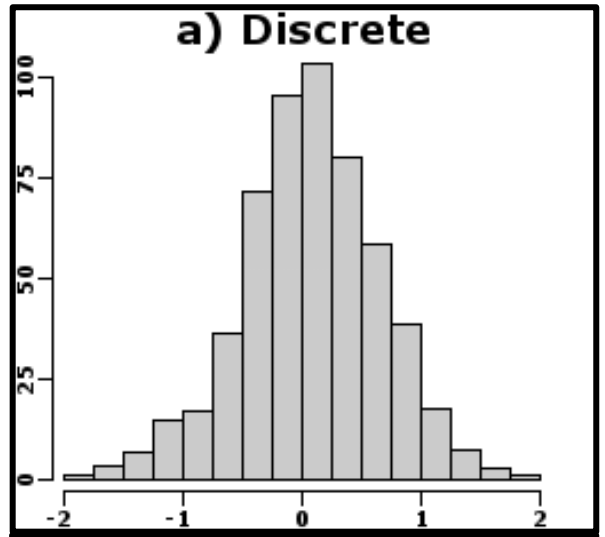
- Gaussian – random dispersal from central point; random diffusion from mean
- Bernoulli Trial – discrete event with only two outcomes (success/failure) with constant probability of success: P(success)= *p*.
- Poisson point process – events that occur individually in continuous time (or space)

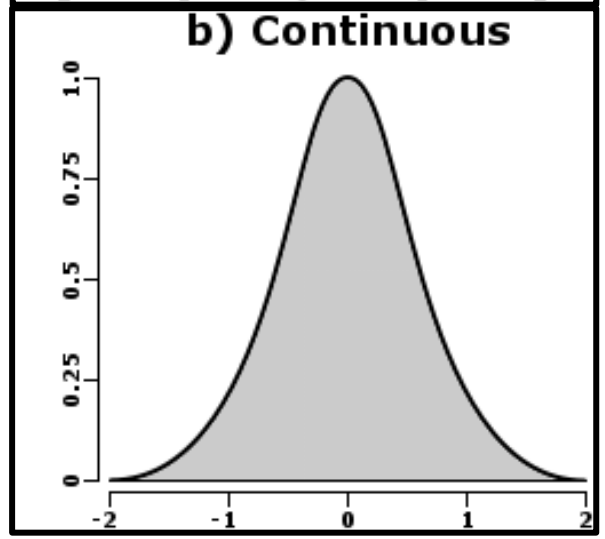
Broad classification

- Discrete - takes on only integer or 'count' variables
- Continuous – takes on any numerical value within range









DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Properties

UND UNIVERSITY OF NORTH DAKOTA.

Name	Notation	Range	Mean(x)	Var(x)	Explanation
Normal	$N(\mu,\sigma^2)$	$-\infty < x < \infty$	μ	σ^2	x=dispersal from a central point, or a diffusion through a Gaussian filter, with variance independent of mean.
Log-Normal	$Lognormal(\mu,\sigma^2)$	$x > 0$	$\exp(\mu+\sigma^2/2)$	$[\exp(\sigma^2)-1]^* \exp(2\mu+\sigma^2)$	x =probability distribution whose logarithm is normally distributed.
Exponential	$exp(\beta)$	$x > 0$	β	β^2	x =time between events that occur at a rate of $\lambda = 1/\beta$.
Gamma	$Gamma(k,\theta)$	$x > 0$	$k\theta$	$k\theta^2$	x =time it takes for *k* events to occur within a rate of $\lambda = 1/\theta$, or the sum of *k* exponential events.
Beta	$Beta(a,b)$	$0 < x < 1$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2+(a+b+1)}$	x =distribution of probabilities based on *a* successes and *b* failures, where both *a* and *b* >1.
Binomial	$Bin(n,p)$	x = 0, 1, 2…	np	$np(1-p)$	x =number of positive events out of n trials each with a probability of success *p*.
Geometric	$G(p)$	x = 1, 2, 3…	$\frac{1}{p}$	$\frac{1-p}{p^2}$	x =number of trials, with probability of success *p*, that are needed to obtain one success.
Negative Binomial	$NB(n,p)$	x = 0, 1, 2…	$\frac{k(1-p)}{p}$	$\frac{k(1-p)}{p^2}$	x =number of failures before *k* successes occur in sequential independent trials, all with the same probability of success, *p*.
Poisson	$Poisson(\lambda)$	x = 0, 1, 2…	λ	λ	x =count of items in a standardized unit of effort that occur at rate λ.

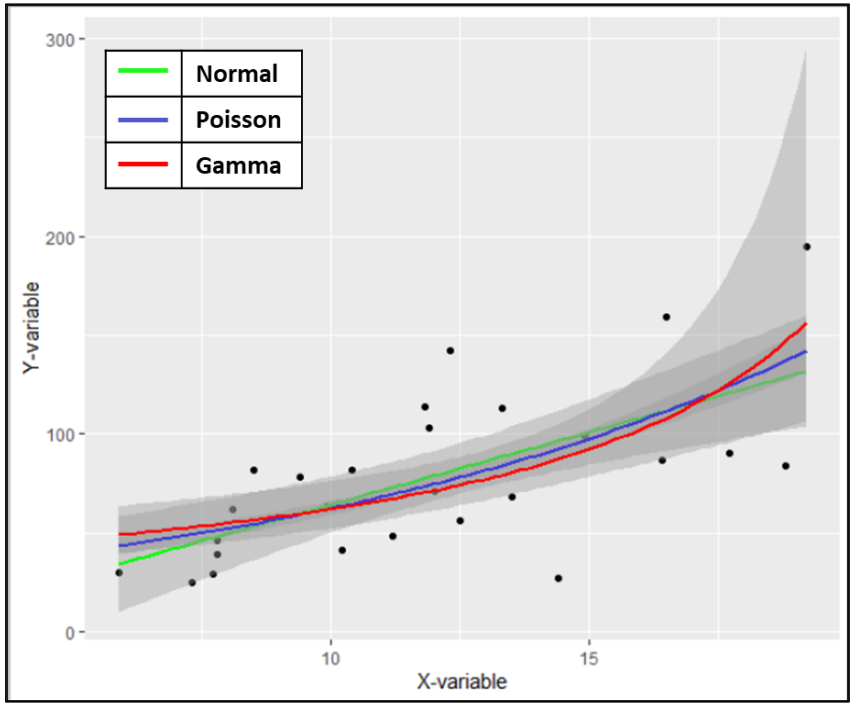
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Statistical tests

UNIVERSITY OF NORTH DAKOTA

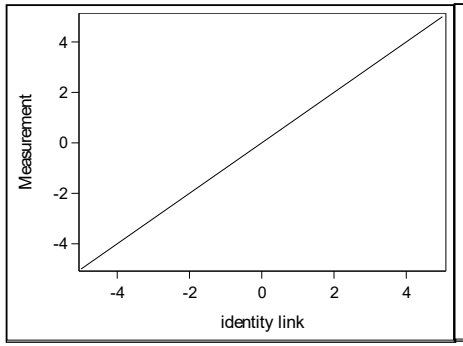
What sort of tests use distributions?

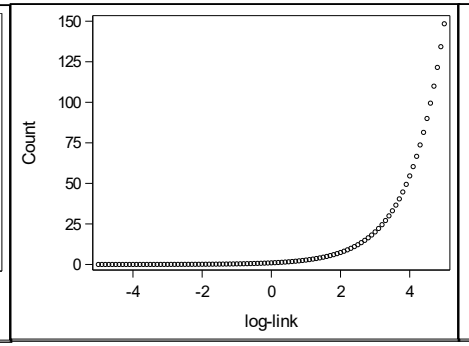
- Parametric tests assume a normal (Gaussian) distribution
  - ANOVA, t-test, linear regression, etc.
- Generalized linear models can use various distributions
  - 'generalized' refers to models that don't assume a normal distribution
  - Common Examples are Poisson and Logistic regression

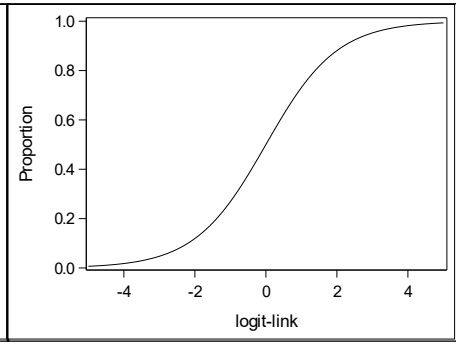


How to plot results?

- Each distribution has a standard link function
- Common ones are 'identity', 'log' and 'logit'







DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

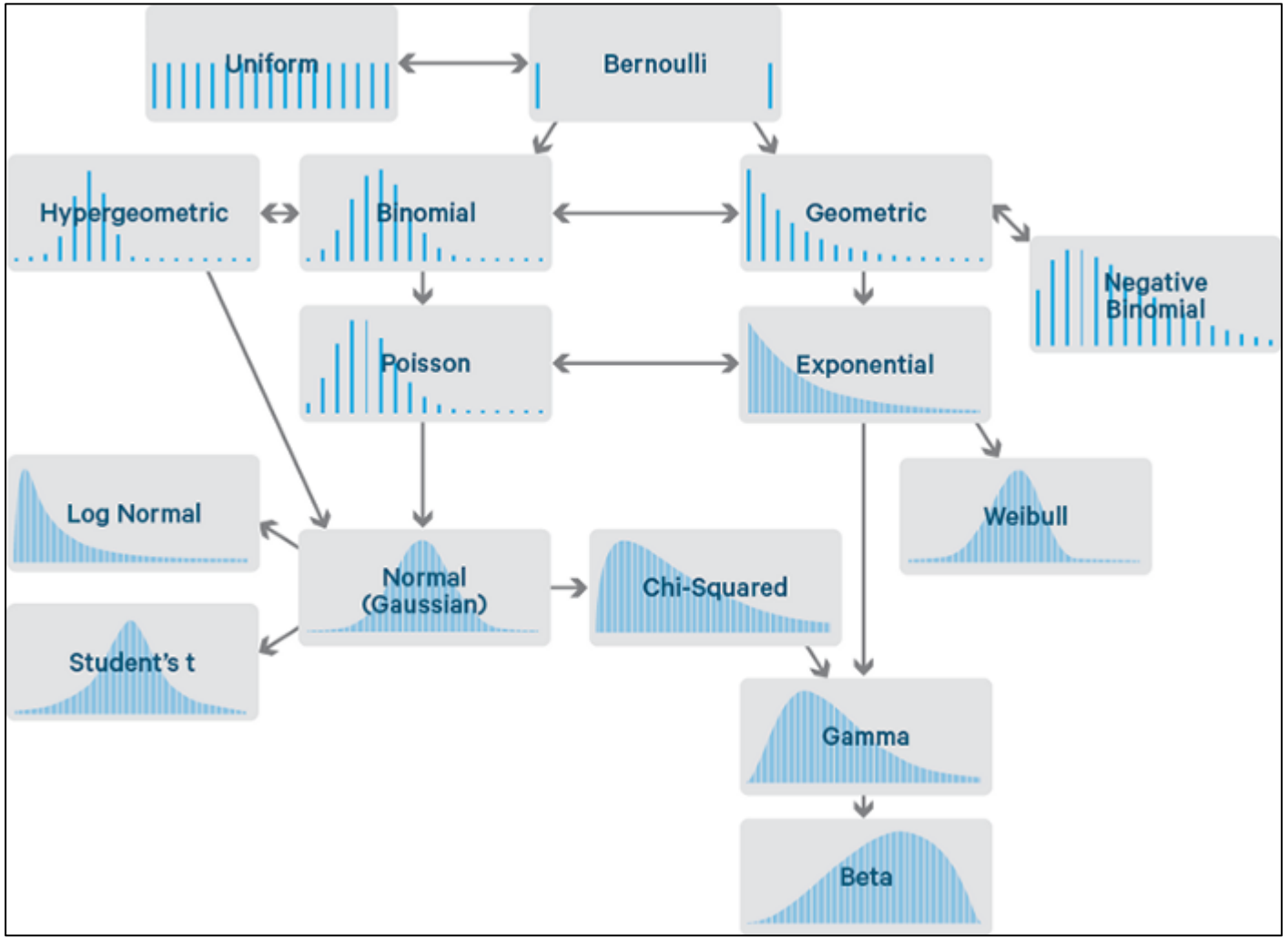
Taxonomy

UNIVERSITY OF NORTH DAKOTA

Taxonomy of Random Variables

Below is a taxonomy of common distributions, classified primarily by the topology of their support.

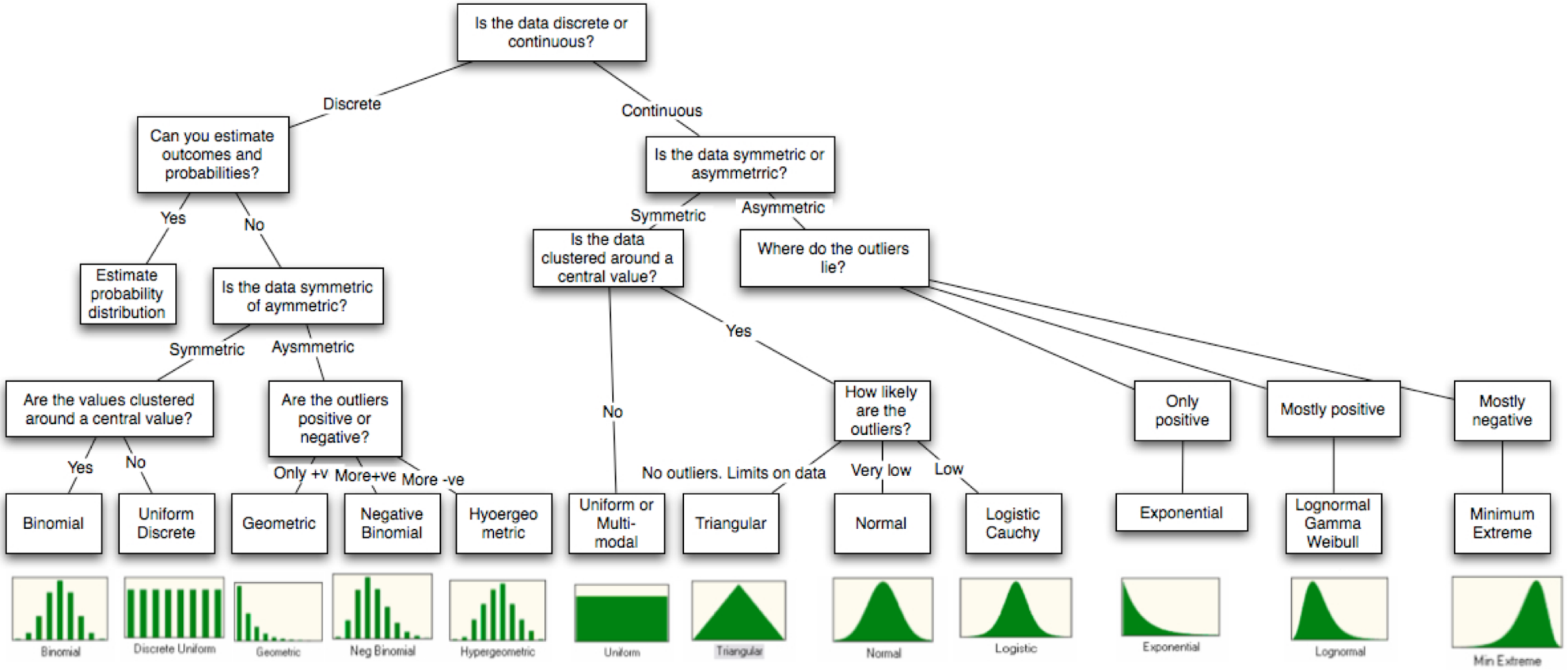
- finite
  - Dirac
  - Bernoulli
- countable
  - binomial
  - geometric
  - Poisson
- interval
  - uniform
  - beta
- half-line
  - exponential
  - gamma
- unbounded
  - normal
  - Cauchy
  - (Lévy) stable
- *transforms*
  - (generalized) Pareto
  - inverse gamma
  - lognormal
- *mixtures*
  - (Gosset) Student t
  - negative-binomial
- *non-parametric*
  - empirical



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Taxonomy cont.

UND UNIVERSITY OF NORTH DAKOTA.



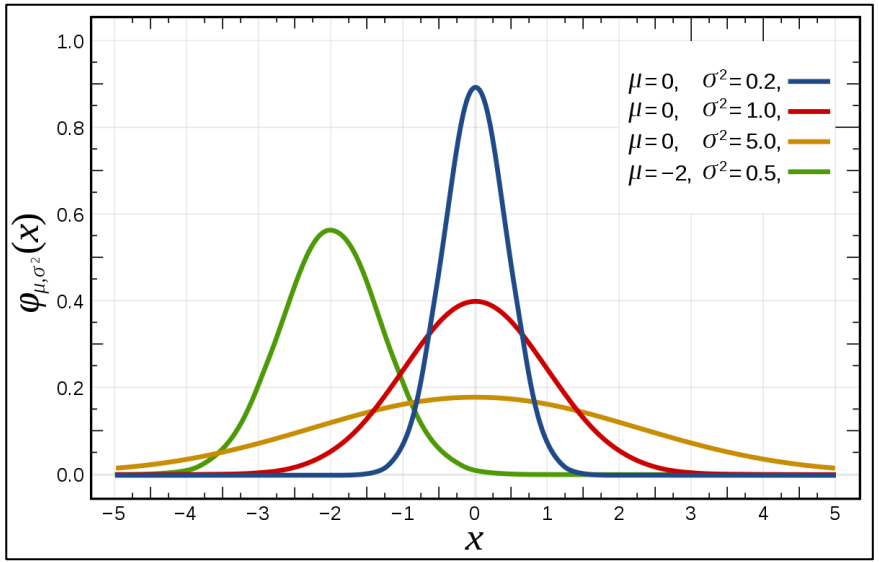
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Normal & Log Normal

UNIVERSITY OF NORTH DAKOTA

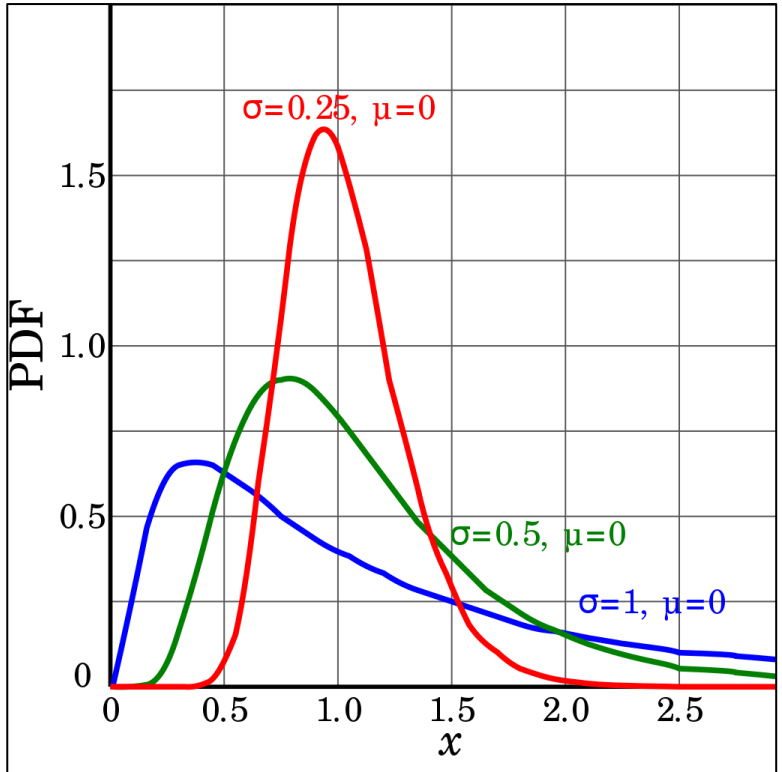
Normal

- Distribution models dispersion from central point
- Continuous, unbounded, and with a variance independent of its mean



Log Normal

- Distribution whose log-transformation is normal



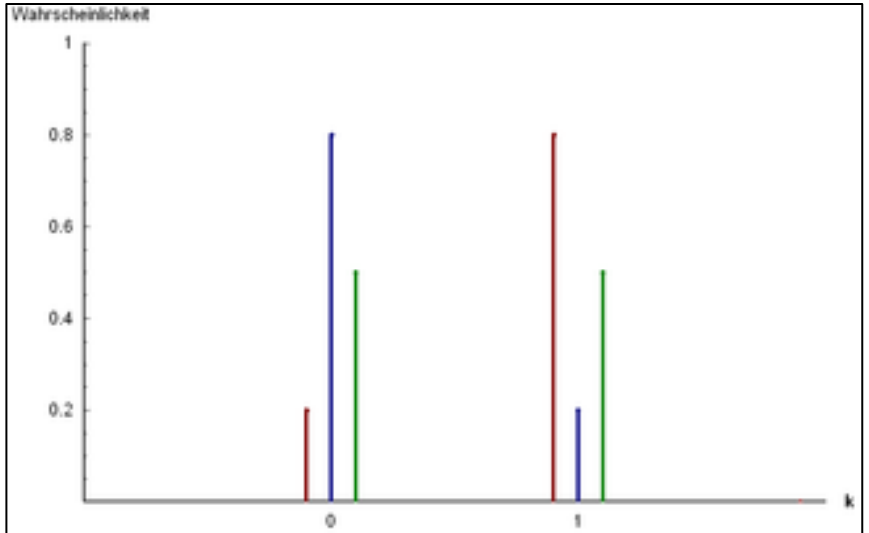
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Binary & Binomial

UND UNIVERSITY OF NORTH DAKOTA.

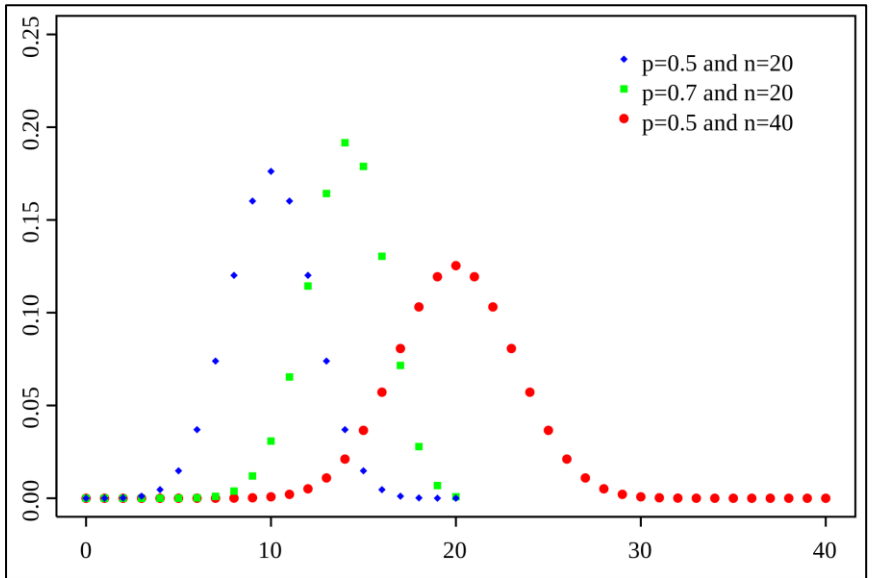
Binary

- Distribution that models binary (two) outcomes
- Aka ‘Bernoulli trials’
- Example: coin flips



Binomial

- Distribution that models the number of success from Bernoulli trials
- Example: coin flips



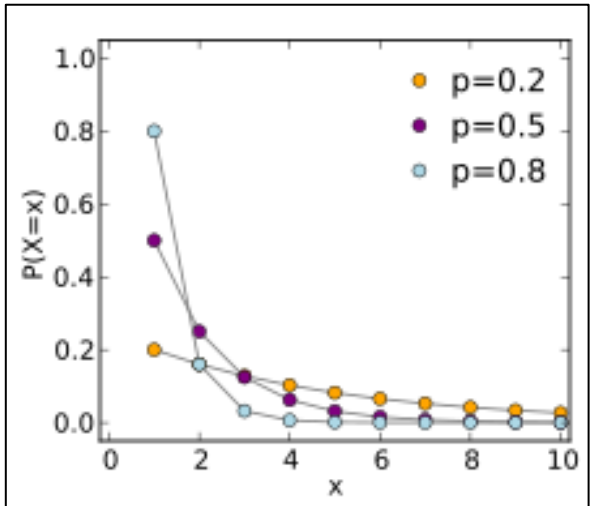
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Geometric, Poisson, & Negative Binomial

UNIVERSITY OF NORTH DAKOTA

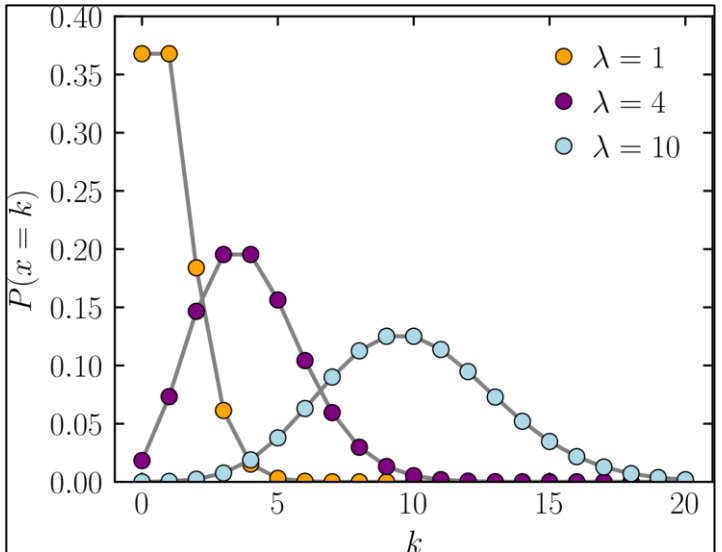
Geometric

- Distribution that models number of Bernoulli trials needed for one success
- Start at 0 or 1
- Example: coin flips until heads



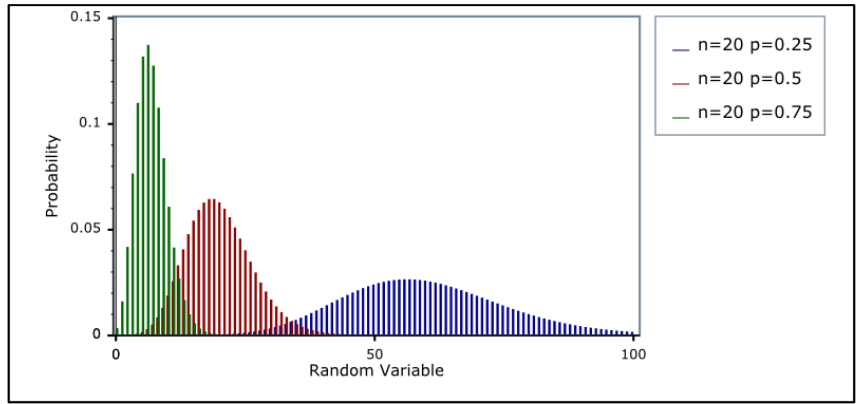
Poisson

- Distribution that models the number of counts occurring at a certain rate
- Example: number of patrons entering restaurant an hour, or number of trees per acre of woods



Negative Binomial

- Distribution that models the number of failures needed for a certain number of successes
- Can also be used for 'count' data as a sort of over-dispersed correction for Poisson
- Example: flip coins until 3 heads, count # of tails



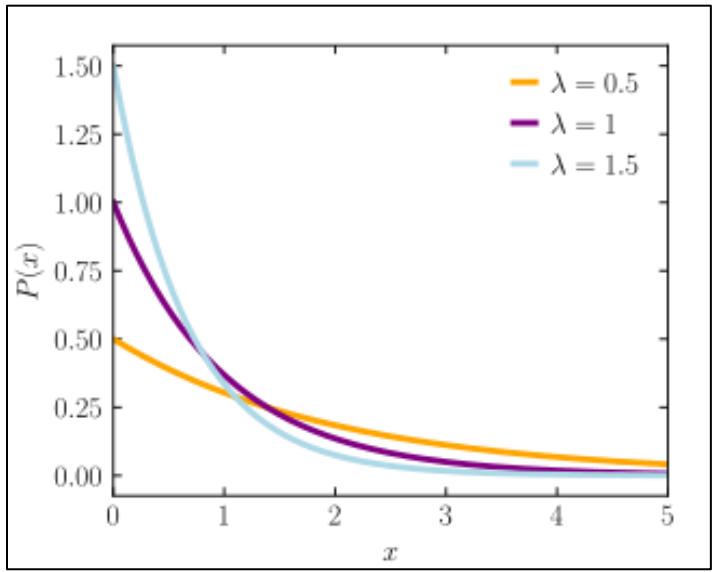
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Exponential, Gamma, & Beta

UND UNIVERSITY OF NORTH DAKOTA.

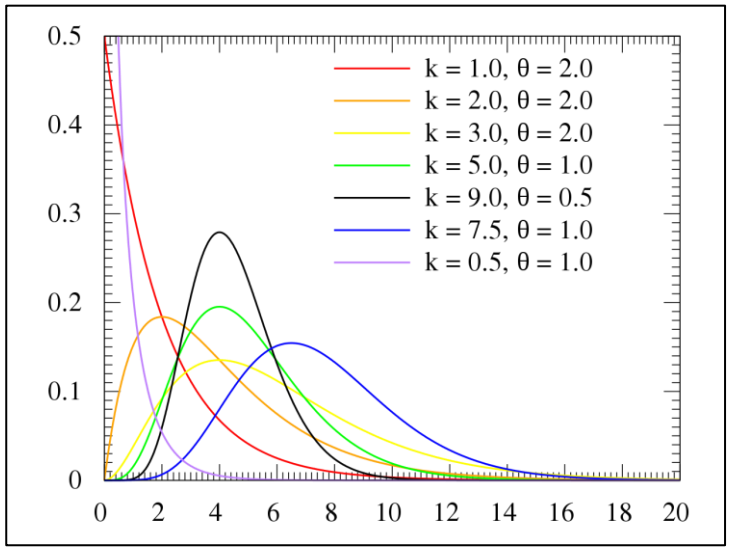
Exponential

- Distribution that models the inter-arrival time between Poisson process events
- Example: time between arrive of patrons into a restaurant



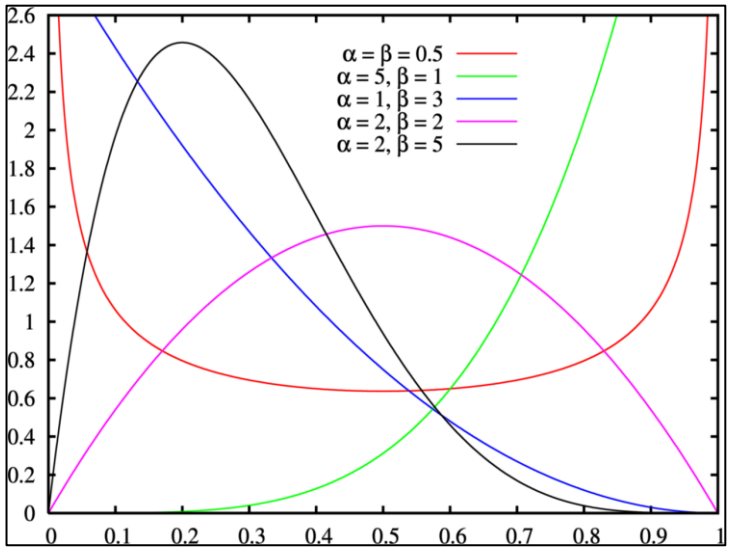
Gamma

- Distribution that models the total arrival time for a number of Poisson process events
- Example: time it takes for 5 patrons to enter a restaurant



Beta

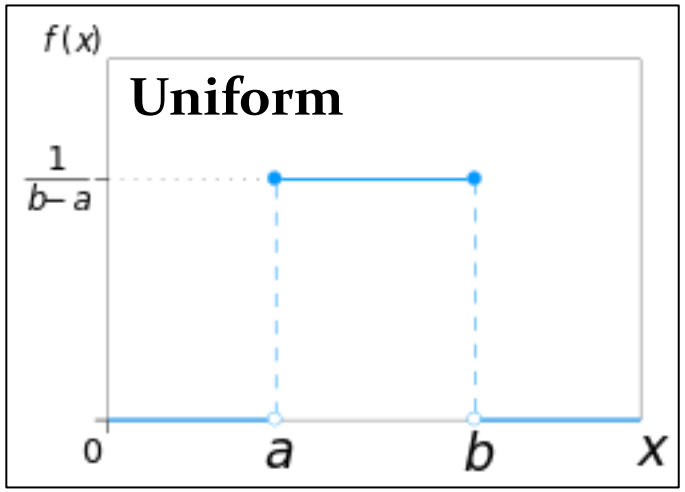
- Distribution that models a proportion of success between 0 and 1
- Example: proportion of hits out of at-bats

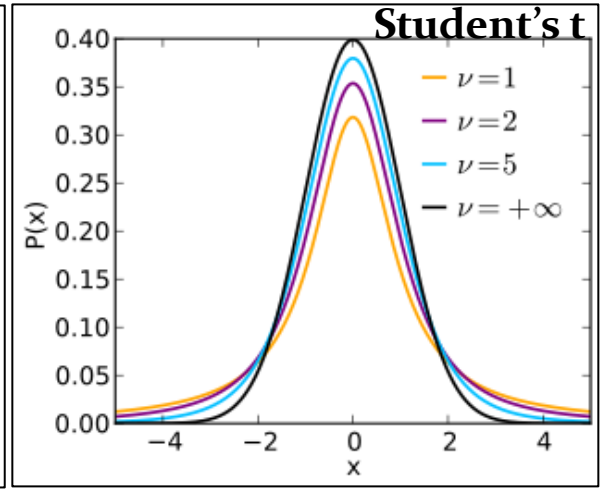


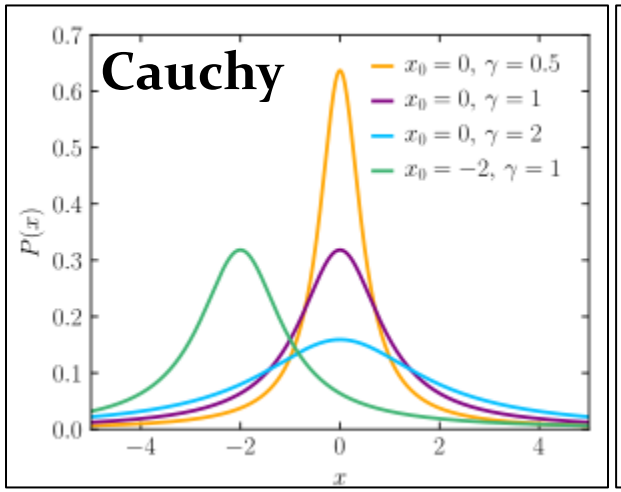
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

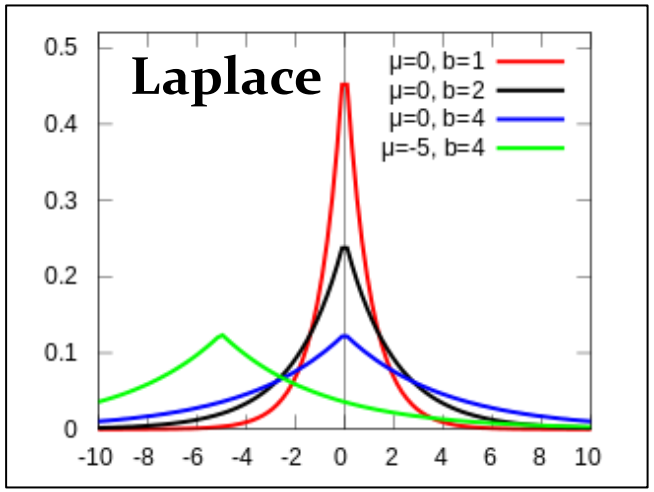
Other distributions

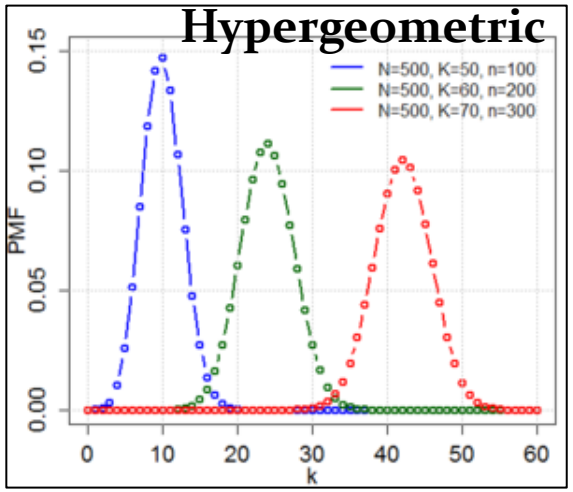
UND UNIVERSITY OF NORTH DAKOTA

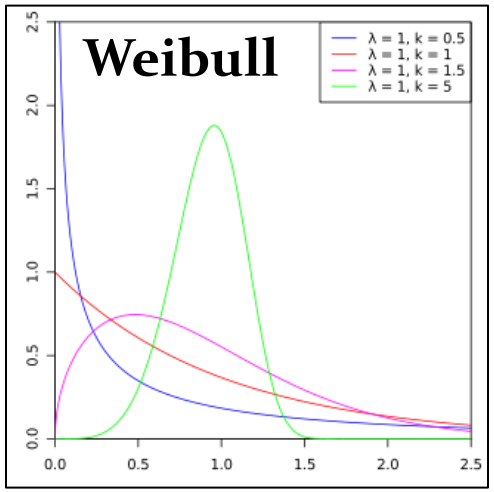


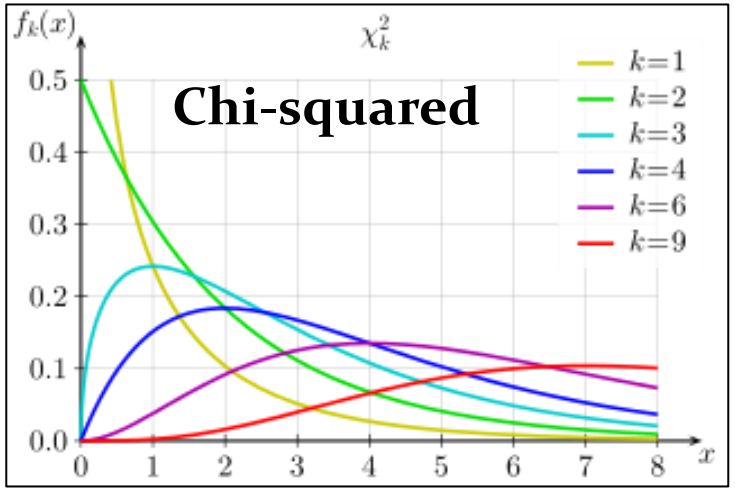


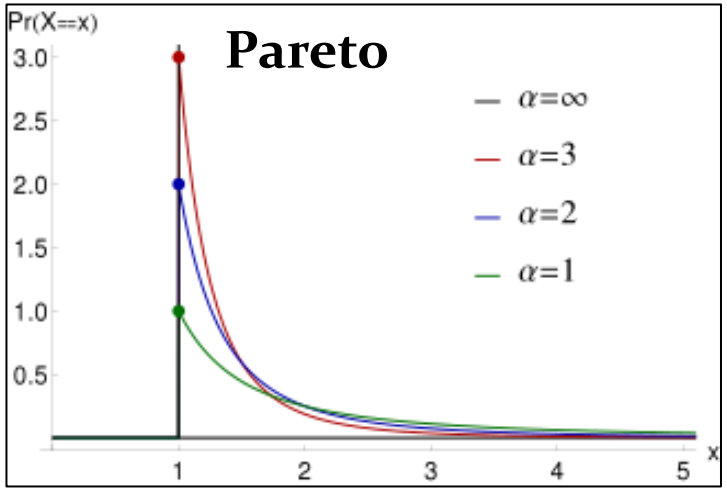












DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

SAS examples

UNIVERSITY OF NORTH DAKOTA.

DIST=	Distribution	Default Link Function	Numeric Value
BETA	beta	logit	12
BINARY	binary	logit	4
BINOMIAL | BIN | B	binomial	logit	3
EXPONENTIAL | EXPO	exponential	log	9
GAMMA | GAM	gamma	log	5
GAUSSIAN | G | NORMAL | N	normal	identity	1
GEOMETRIC | GEOM	geometric	log	8
INVGAUSS | IGAUSSIAN | IG	inverse Gaussian	inverse squared	6
		(power(–2))	
LOGNORMAL | LOGN	lognormal	identity	11
MULTINOMIAL | MULTI | MULT	multinomial	cumulative logit	NA
NEGBINOMIAL | NEGBIN | NB	negative binomial	log	7
POISSON | POI | P	Poisson	log	2
TCENTRAL | TDIST | T	t	identity	10
BYOBS(*variable*)	multivariate	varied	NA

SAS Programmer | Sign Out

Distributions Examples.sas

CODE | LOG | RESULTS

```
*Create and view distributions;
DATA rand;
    do i=1 to 10000;
        * generate random values for 10000 observations;
        trt="trt";
        mynormal=rand("NORMAL", 85, 40);
        mylognormal=rand("LOGNORMAL", 1, 0.1);
        mybeta=rand("BETA", 1,99);
        mybinary=rand("BINOM", 0.95, 1);
        mybinomial=rand("BINOM", 0.95, 2000);
        myexponential=rand("EXPONENTIAL", 4.2);
        mygamma=rand("GAMMA", 2, 2);
        mygeometric=rand("GEOMETRIC", 0.10);
        mynegbinomial=rand("NEGBINOMIAL", 0.97, 1000);
        mypoisson=rand("POISSON", 10);
        output;
    end;

PROC PRINT data=rand(obs=20);

PROC UNIVARIATE data=rand;
    var mynormal mylognormal mybeta mybinary mybinomial myexponential mygamma mygeometric
        mynegbinomial mypoisson;
    histogram;

    *can change around the parameters;
*-------------------------------------------------------------------------------------------------------------------------------;

*Examples of distributions in action;

    *Datasets;
DATA Fish; set sashelp.Fish;
PROC PRINT data=Fish(obs=25);

DATA Baseball; set sashelp.Baseball;
PROC PRINT data=Baseball(obs=25);


    *Normal;
PROC UNIVARIATE data=Fish;
```

/home/markwilliamson20/my_courses/markwilliamson0/MW_2021_Work/Module Examples/Distributions Examples.sas UTF-8

Messages | User: markwilliamson20

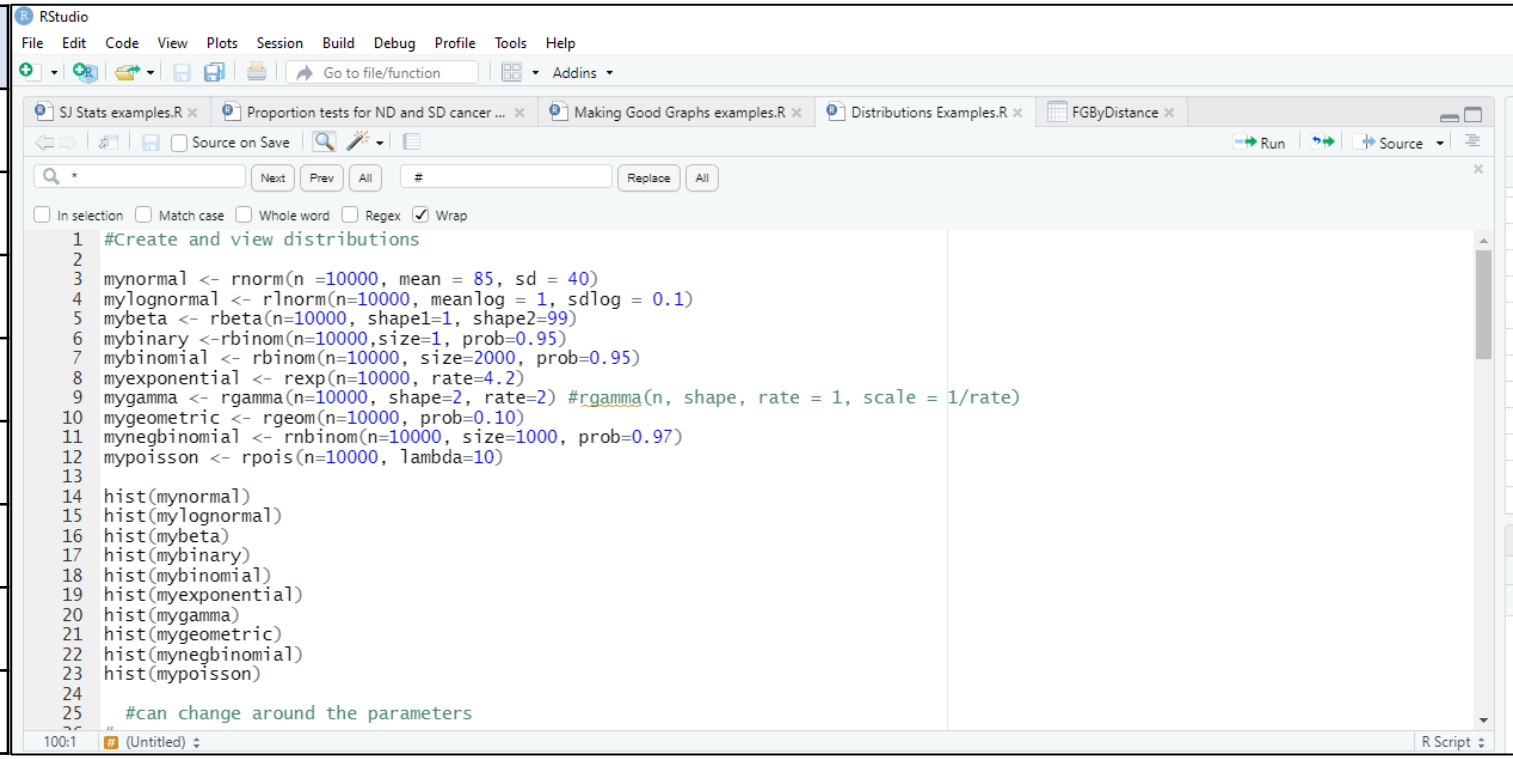
Code available at: https://med.und.edu/daccota/_files/docs/berdc_docs/distributions_sas_code.txt

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

R examples

UND UNIVERSITY OF NORTH DAKOTA.

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "logit")

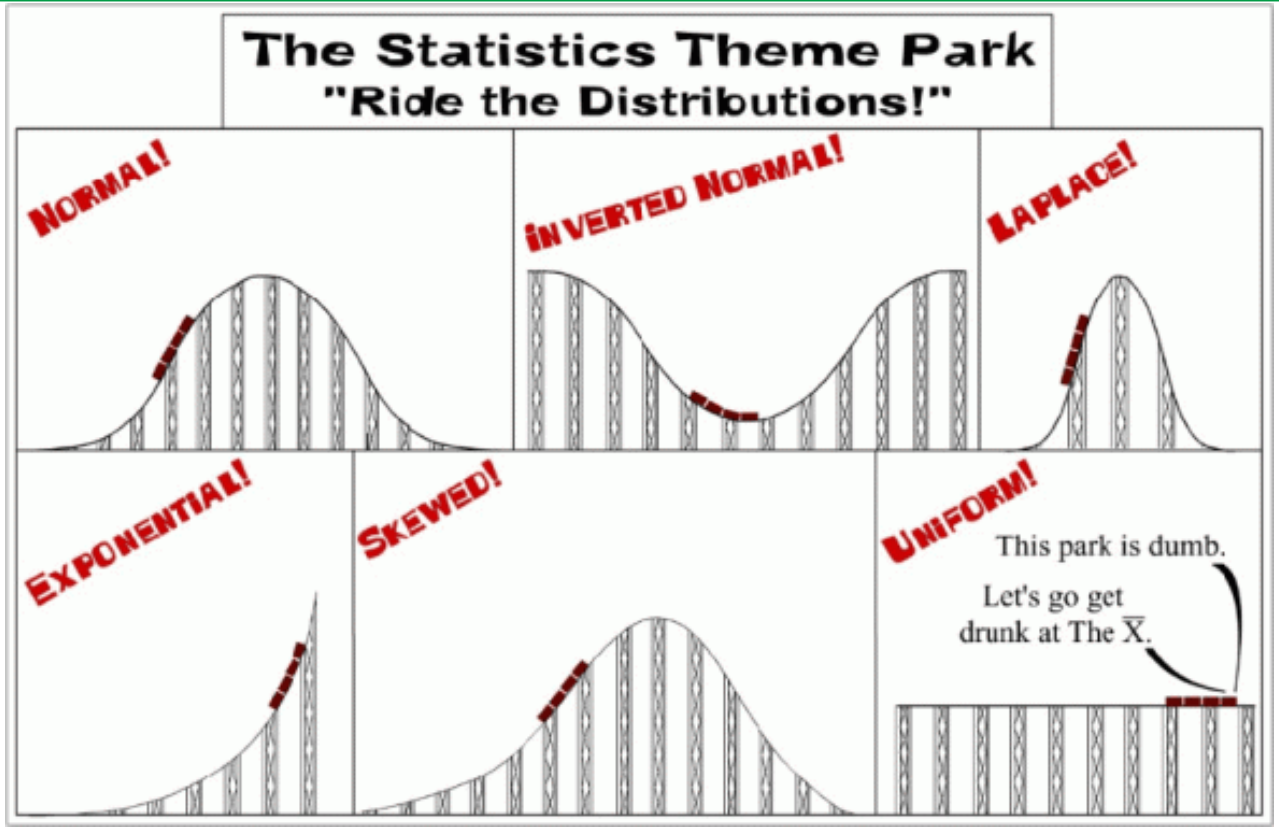


Code available at: https://med.und.edu/daccota/_files/docs/berdc_docs/distributions_r_code.txt

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Wrap-up



Please take the post-test and survey:

Post-test: https://und.qualtrics.com/jfe/form/SV_5swKhmpU4tO7hau

Survey: https://und.qualtrics.com/jfe/form/SV_5o370oBxAOTtMr4

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

References

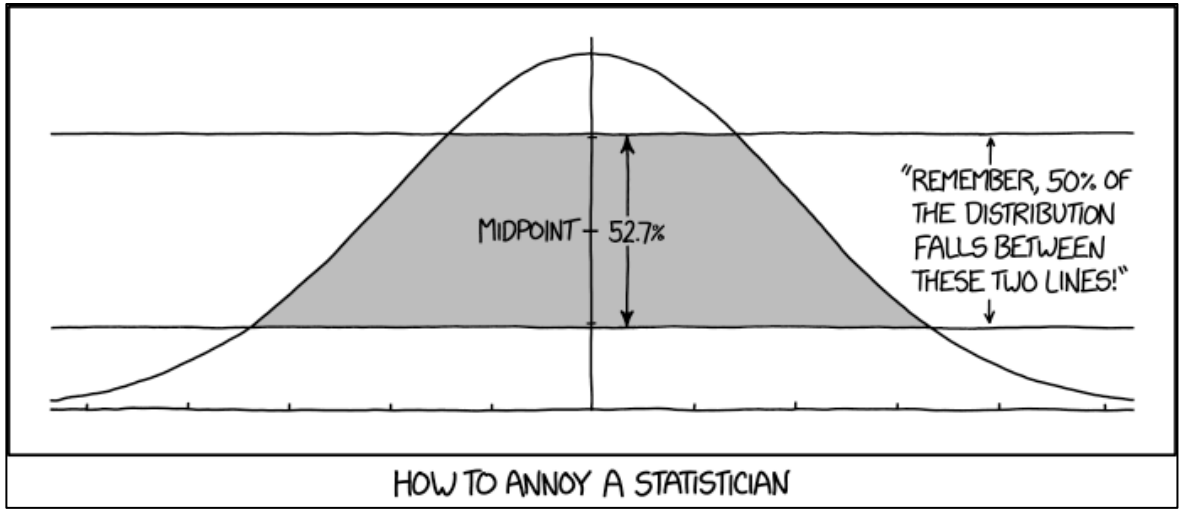
UNIVERSITY OF NORTH DAKOTA

Images:

- https://www.wolfram.com/mathematica/newin6/content/SymbolicStatisticalComputing/BuiltinStatisticalDistributions.html
- https://pierpaolo28.github.io/blog/blog19/
- https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/continuous-probability-distribution-2-of-2/
- https://www.spss-tutorials.com/sampling-distribution-what-is-it/
- https://medium.com/analytics-vidhya/probability-distributions-444e7babf2e1
- https://www.kdnuggets.com/2020/02/probability-distributions-data-science.html
- https://www-users.cse.umn.edu/~dodso013/fm503/0910/lectures/fall2.pdf
- https://tinyheero.github.io/2016/03/17/prob-distr.html
- https://en.wikipedia.org/wiki/List_of_probability_distributions
- https://loonylabs.files.wordpress.com/2019/10/distribution-park-joke.gif?w=590

Materials:

- https://365datascience.com/tutorials/statistics-tutorials/distribution-in-statistics/
- https://www.itl.nist.gov/div898/handbook/eda/section3/eda361.htm
- https://www.statisticshowto.com/negative-binomial-experiment/
- https://www-users.cse.umn.edu/~dodso013/fm503/0910/lectures/fall2.pdf
- https://medium.com/@srowen/common-probability-distributions-347e6b945ce4
- https://www.causascientia.org/math_stat/Dists/Compendium.pdf



https://xkcd.com/2118/

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Acknowledgements

- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.