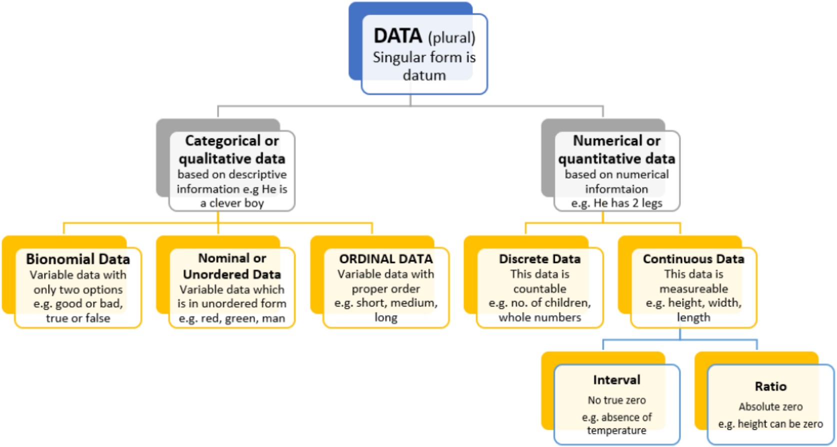
Stats for Data Science

Saumya Bhatnagar

February 9, 2020

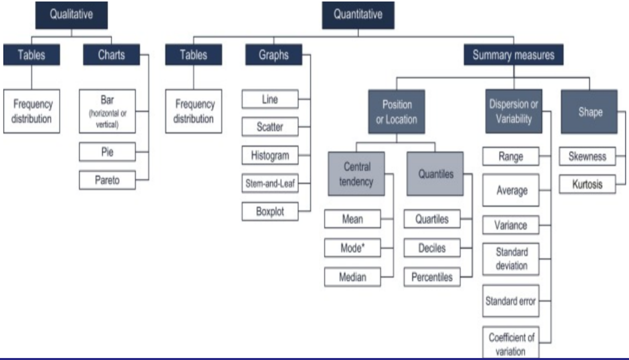
Table of contents

Initial Terminologies



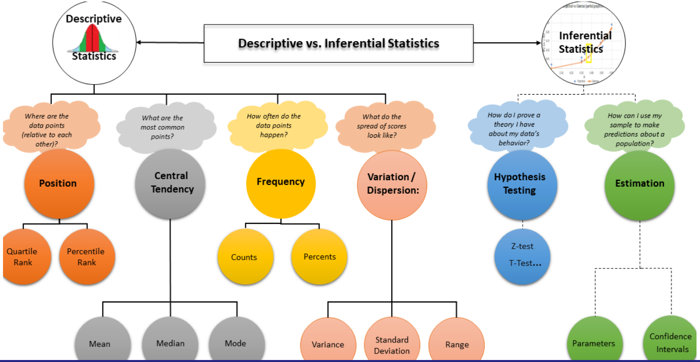
Types of Analysis

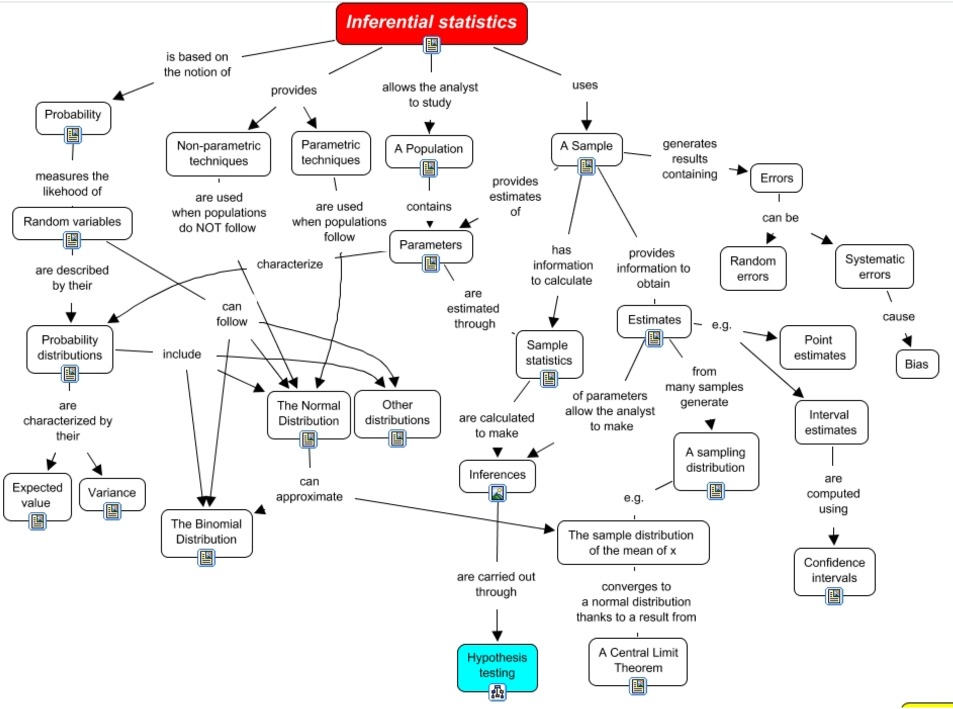
- Qualitative Analysis/Non-Statistical Analysis gives generic information (uses text, sound and other forms of media).
- Quantitative Analysis/Statistical Analysis: collecting and interpreting data.



Types of Statistics

- Descriptive Statistics: provides descriptions of the population.
- Inferential Statistics makes inferences and predictions from sample to generalize a population.





Inferential statistics
is based on the notion of
provides
allows the analyst to study
uses
Probability
Non-parametric techniques
Parametric techniques
A Population
A Sample
generates results containing
Errors
measures the likehood of
are used when populations do NOT follow
are used when populations follow
contains
provides estimates of
can be
Random variables
Parameters
has information to calculate
provides information to obtain
Random errors
Systematic errors
characterize
are described by their
can follow
are estimated through
Estimates
e.g.
Point estimates
cause
Probability distributions
include
Sample statistics
from many samples generate
Bias
The Normal Distribution
Other distributions
of parameters allow the analyst to make
Interval estimates
are characterized by their
are calculated to make
A sampling distribution
Expected value
Variance
Inferences
e.g.
are computed using
can approximate
The Binomial Distribution
The sample distribution of the mean of x
Confidence intervals
are carried out through
converges to a normal distribution thanks to a result from
Hypothesis testing
A Central Limit Theorem

Contingency Table and Probabilities

Joint, Marginal and Conditional

- Joint probabilities for rain and wind:

	no wind	some wind	strong wind	storm
no rain	0.1	0.2	0.05	0.01
light rain	0.05	0.1	0.15	0.04
heavy rain	0.05	0.1	0.1	0.05

- Marginalize to get simple probabilities:
  - $P(\text{no wind}) = 0.1 + 0.05 + 0.05 = 0.2$
  - $P(\text{light rain}) = 0.05 + 0.1 + 0.15 + 0.04 = 0.34$
- Combine to get conditional probabilities:
  - $P(\text{no wind}|\text{light rain}) = \frac{0.05}{0.34} = 0.147$
  - $P(\text{light rain}|\text{no wind}) = \frac{0.05}{0.2} = 0.25$

- Variable and Random Variable (RV)
- Parameter and Hyper-parameter
- Mean, Median, Mode
- mode sucks for small samples
- Range, IQR
- Standard Deviation (σ): Measure of the how spread out data is from its mean.
- Variance (σ^2): It describes how much a random variable differs from its expected value. It entails computing squares of deviations.The average of the squared differences from the Mean.
    1. Deviation is the difference bw each element from the mean.
    2. Population Variance = avg of squared deviations
    3. Sample Variance = avg of squared differences from the mean

EXPECTED VALUE

Discrete random variable $E(X) = \sum_x x\, p_x(x)$

- Provided $\sum_x |x|\, p_x(x) < \infty$. If the sum diverges, the expected value does not exist. **For the jar full of numbered balls**
- A ball is selected at random; all balls are equally likely to be chosen $P(X = x_i) = \frac{1}{N}$.
- Say n_1 balls have value v_1, and n_2 balls have value v_2, and $\ldots n_n$ balls have value v_n. Unique values are v_i, for $i = 1, \ldots, n$. Note $n_1 + \cdots + n_n = N$, and $P(X = v_j) = \frac{n_j}{N}$. $E(X) = \frac{\sum_{i=1}^{N} x_i}{N}$

Continuous random variable $E(X) = \int_{-\infty}^{\infty} x\, f_x(x)\, dx$

- Provided $\int_{-\infty}^{\infty} |x|\, f_x(x)\, dx < \infty$. If the integral diverges, the expected value does not exist.

Sometimes the expected value does not exist

Need $\int_{-\infty}^{\infty} |x| \, f_x(x) \, dx < \infty$

For the Cauchy distribution, $f(x) = \frac{1}{\pi(1+x^2)}$.

$$
\begin{aligned}
E(|X|) &= \int_{-\infty}^{\infty} |x| \, \frac{1}{\pi(1+x^2)} \, dx \\
&= 2 \int_{0}^{\infty} \frac{x}{\pi(1+x^2)} \, dx \\
& \quad u = 1 + x^2, \;\; du = 2x \, dx \\
&= \frac{1}{\pi} \int_{1}^{\infty} \frac{1}{u} \, du \\
&= \ln u \Big|_{1}^{\infty} \\
&= \infty - 0 = \infty
\end{aligned}
$$

=> an integral “equals” infinity, it is unbounded above.

For a RV X with PDF $\rho(x)$. The variance($\mathbb{V}$) and the standard deviation(σ_X) of X, are defined by
Variance $\sigma^2 = (1/n)\sum_{i=1}^{n}(x_i - \mu)^2$

$$\mathbb{V} = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_D (x - \mathbb{E})^2 \, dP.$$

$$\mathbb{V} = \int_D x^2 \, dP - \mathbb{E}^2.$$

$$\sigma_X = \sqrt{V[X]} \qquad = \sqrt{E[X^2] - E[X]^2}$$

$$\mathbb{V} = \sqrt{\int_D x^2\rho(x)\,dx - \left(\int_D x\rho(x)\,dx\right)^2}.$$

If one interprets the PDF ($\rho(x)$) as the density of a rod at location (x), then:
The mean, ($\mu = \int x\rho(x)\,dx$), gives the center of mass of the rod.
The variance, ($V = \int (x-\mu)^2\rho(x)\,dx$), gives the moment of inertia about the line ($x = \mu$).
The standard deviation, ($\sigma = \sqrt{V}$), gives the radius of gyration about the line ($x = \mu$).

Std error vs std deviation

$$\text{std error} =$$

coeff of variation

$CV = \frac{sd}{\bar{x}}, where \bar{x} = samplemean$
$x = [1, 2, 3] => \bar{x} = 2$ and $S_x = 1 => CV(x) = 1/2$
$y = [101, 102, 103] => \bar{y} = 102$ and $S_y = 1 => CV(y) = 1/102$
Higher the CV means higher fluctuations in the dataset

skewness and kurtosis

skewness

mode skewness $= \frac{mean - mode}{stddev}$

in skewed data: mode = 3(median) - 2(mean)

for small dataset, use below:

median skewness $= \frac{3(mean - median)}{stddev}$

$$skewness = \begin{cases} approx\_symmetric, & -0.5 <= x <= 0.5 \\ moderately\_skewed, & 0.5 < |x| < 1 \\ highly\_skewed, & |x| > 1 \end{cases} \tag{1}$$

kurtosis: same mean or sd but diff peakedness

higher peaked => higher kurtosis

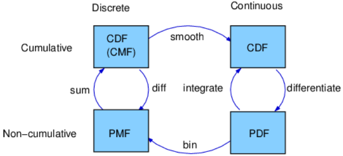
Moments

I moment: $\frac{\sum x}{n}$ => **mean** => considered as values from 0
second moment: $\frac{\sum x^2}{n}$ =>values further from 0 will be higher, so instead we take centralized
second (centralized) moment: $\frac{\sum (x-\mu)^2}{n}$ => variance
third (centralized) moment: $\frac{1}{n}\frac{\sum (x-\mu)^3}{\sigma^3}$ − > skew
but since we don't have population mean, we have sample mean, we adjust the above value with degrees of freedom
II (centralized) moment: $\frac{\sum (x-\bar{x})^2}{n-1}$ => **variance**
III (centralized) moment: $\frac{n}{(n-1)(n-2)}\frac{\sum (x-\bar{x})^3}{s^3}$ => **skew**
IV moment: $\frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{\sum (x-\bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$ => **kurtosis**

Few examples of distributions are,

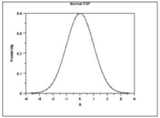
- Discrete
  1. Uniform Discrete distribution or Rectangular Dist
  2. Geometric disctribution
  3. Binominal distribution
  4. Poission distribution
- Continuous
  1. Uniform distribution
  2. Normal distribution/Gaussian distribution/Bell Curve
  3. Student's T distribution... to check
  4. Gamma distribution
  5. Exponential distribution
  6. Bernoulli distribution ... to check
  7. Beta
  8. Triangular

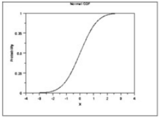
- A PMF, "f" returns the probability of an outcome: $f(x) = P(X = x)$

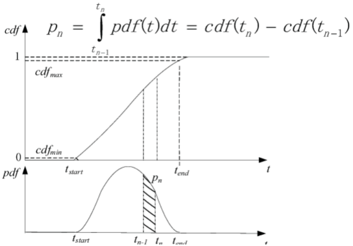


- Reliability function & Hazard Function

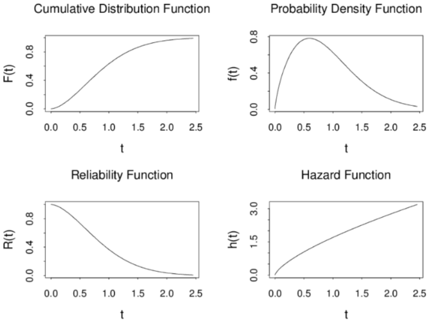
- Probability density function (p.d.f.) denote f
- Probability mass function (p.m.f.) denote f
  f(x)=P(X=x)
- Cumulative distribution function (c.d.f.) denote F
  F(x)=P(X≤x)







- Reliability function & Hazard Function



Types of Distributions

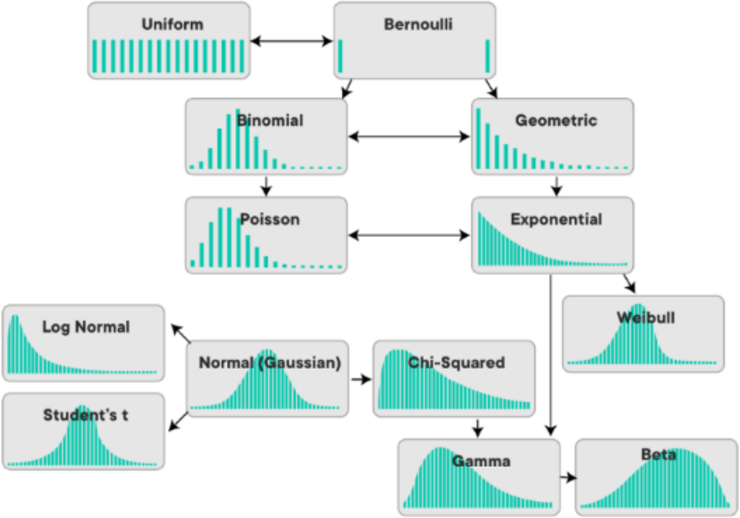
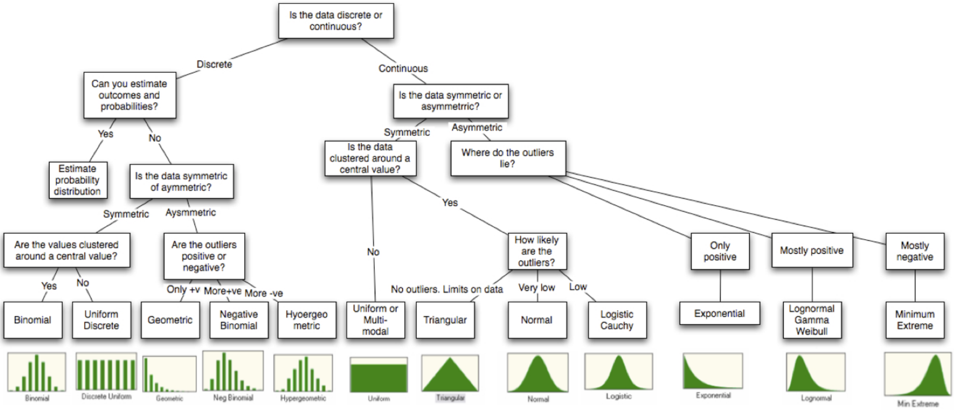


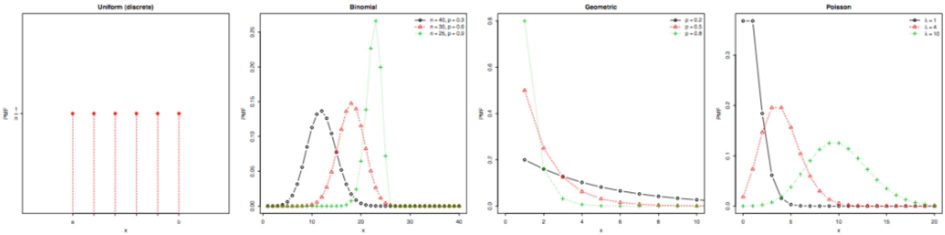
Figure 6A.15: Distributional Choices



1.1 Discrete Distributions

CDF/CMF PMF Expected Val of RV Var of RV

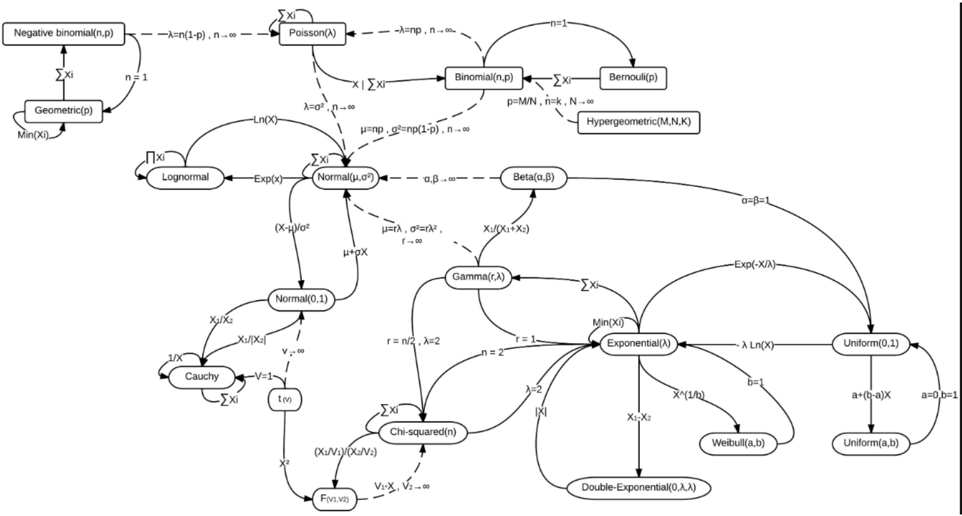
	Notation[1]	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
Uniform	$\text{Unif}\{a,\dots,b\}$	$\begin{cases} 0 & x<a \\ \frac{\lfloor x\rfloor - a+1}{b-a} & a \le x \le b \\ 1 & x>b \end{cases}$	$\frac{I(a<x<b)}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{e^{as}-e^{-(b+1)s}}{s(b-a)}$
Bernoulli	$\text{Bern}(p)$	$(1-p)^{1-x}$	$p^x(1-p)^{1-x}$	p	$p(1-p)$	$1-p+pe^s$
Binomial	$\text{Bin}(n,p)$	$I_{1-p}(n-x,x+1)$	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$	$(1-p+pe^s)^n$
Multinomial	$\text{Mult}(n,p)$		$\frac{n!}{x_1!\dots x_k!}p_1^{x_1}\cdots p_k^{x_k} \quad \sum_{i=1}^{k} x_i = n$	np_i	$np_i(1-p_i)$	$\left(\sum_{i=0}^{k} p_i e^{s_i}\right)^n$
Hypergeometric	$\text{Hyp}(N,m,n)$	$\approx \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right)$	$\frac{\binom{m}{x}\binom{m-x}{n-x}}{\binom{N}{x}}$	$\frac{nm}{N}$	$\frac{nm(N-n)(N-m)}{N^2(N-1)}$	
Negative Binomial	$\text{NBin}(n,p)$	$I_p(r,x+1)$	$\binom{x+r-1}{r-1}p^r(1-p)^x$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$	$\left(\frac{p}{1-(1-p)e^s}\right)^r$
Geometric	$\text{Geo}(p)$	$1-(1-p)^x \quad x\in\mathbb{N}^+$	$p(1-p)^{x-1} \quad x\in\mathbb{N}^+$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{p}{1-(1-p)e^s}$
Poisson	$\text{Po}(\lambda)$	$e^{-\lambda}\sum_{i=0}^{x}\frac{\lambda^i}{i!}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	$e^{\lambda(e^s-1)}$



1.2 Continuous Distributions

	Notation	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
Uniform	$\text{Unif}(a,b)$	$\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$	$\frac{I(a < x < b)}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{sb}-e^{sa}}{s(b-a)}$
Normal	$\mathcal{N}(\mu,\sigma^2)$	$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt$	$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$\exp\left\{\mu s + \frac{\sigma^2 s^2}{2}\right\}$
Log-Normal	$\ln\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{\ln x - \mu}{\sqrt{2\sigma^2}}\right]$	$\frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$	
Multivariate Normal	$\text{MVN}(\mu,\Sigma)$		$(2\pi)^{-k/2}\|\Sigma\|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$	μ	Σ	$\exp\left\{\mu^T s + \frac{1}{2}s^T\Sigma s\right\}$
Student's t	$\text{Student}(\nu)$	$I_x\left(\frac{\nu}{2},\frac{\nu}{2}\right)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	0	0	
Chi-square	χ^2_k	$\frac{1}{\Gamma(k/2)}\gamma\left(\frac{k}{2},\frac{x}{2}\right)$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2} e^{-x/2}$	k	$2k$	$(1-2s)^{-k/2}\ s < 1/2$
F	$\text{F}(d_1,d_2)$	$I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2},\frac{d_1}{2}\right)$	$\frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\text{B}\left(\frac{d_1}{2},\frac{d_1}{2}\right)}$	$\frac{d_2}{d_2-2}$	$\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$	
Exponential	$\text{Exp}(\beta)$	$1-e^{-x/\beta}$	$\frac{1}{\beta}e^{-x/\beta}$	β	β^2	$\frac{1}{1-\beta s}\ (s < 1/\beta)$
Gamma	$\text{Gamma}(\alpha,\beta)$	$\frac{\gamma(\alpha, x/\beta)}{\Gamma(\alpha)}$	$\frac{1}{\Gamma(\alpha)\,\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	$\alpha\beta$	$\alpha\beta^2$	$\left(\frac{1}{1-\beta s}\right)^\alpha\ (s < 1/\beta)$
Inverse Gamma	$\text{InvGamma}(\alpha,\beta)$	$\frac{\Gamma\left(\alpha,\frac{\beta}{x}\right)}{\Gamma(\alpha)}$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\frac{\beta}{\alpha-1}\ \alpha > 1$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)^2}\ \alpha > 2$	$\frac{2(-\beta s)^{\alpha/2}}{\Gamma(\alpha)} K_\alpha\left(\sqrt{-4\beta s}\right)$
Dirichlet	$\text{Dir}(\alpha)$		$\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$	$\frac{\alpha_i}{\sum_{i=1}^k \alpha_i}$	$\frac{\mathbb{E}[X_i](1-\mathbb{E}[X_i])}{\sum_{i=1}^k \alpha_i + 1}$	
Beta	$\text{Beta}(\alpha,\beta)$	$I_x(\alpha,\beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1+\sum^{\infty}\left(\prod^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{s^k}{k!}$

Types of Distributions



standard uniform density

parameters a = 0 and b = 1, so the PDF for standard uniform density is given by:

$$f(x) = \begin{cases} 1, & 0 <= x <= 1 \\ 0, & otherwise \end{cases} \quad (2)$$

Normal Distribution

Standard Normal Distribution $\mu = 0$; $\sigma = 1$
The 68-95-99.7 rule: Given a normally distributed random variable: $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx .68$ =>68% of samples fall within 1 SD of the mean
$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx .95$
$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx .997$
characteristics of Normal distribution:

1. Mean = median = mode
2. The distribution curve is bell-shaped and symmetrical about the line x=μ.
3. The total AUC = 1.
4. Exactly half of the values are to the left of the center and the other half to the right.

Beta distribution

Connections to other distributions

Prior and Posterior

1.

Binomial, neg-binomial, geometric, hyper-geometric

Conjugate prior for binomial $x|p\ Bin(n,p); p\ Beta(a,b)[prior]$
$f(p|X=k) = usebayes\ rel = placewithbetaabdbin$
$p|X\ Beta(a+X, b+n-x)$
the properties of a Binomial Distribution are

1. Each trial is independent.
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is same for all trials. (Trials are identical.)

Poisson

Characteristics of Poisson distribution:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

λ is the rate at which an event occurs, t is the length of a time interval, And X is the number of events in that time interval. X is called a Poisson RV

Let μ denote the mean number of events in an interval of length t.

Then, $\mu = \lambda$*t.

The PMF of X: $P(X = x) = e^{-\mu} * \frac{\mu^x}{x!}$

Exponential Dist

rate parameter $= \lambda = 1/\beta$

Memoryless property: $P(X >= s + t | X >= s) = P(X >= t)$

$P(X >= s) = 1 - CDF = 1 - P(X <= s) = e^{-\lambda s}$

$P(X >= s + t | X >= s) = \frac{P(X >= s+t, X >= s)}{P(X >= s)}$

$P(X >= s + t | X >= s) = \frac{P(X >= s+t)}{P(X >= s)}$

$P(X >= s + t | X >= s) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}}$

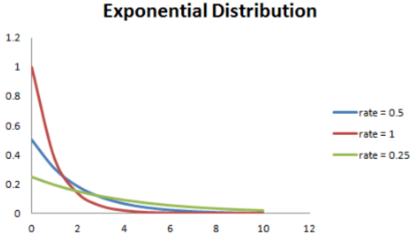
$P(X >= s + t | X >= s) = e^{-\lambda t}$

Memoryless property:

$E(X | X > a) = a + E(X - a | X > a)$

$E(X | X > a) = a + 1/\lambda$

failure rate of any device at time t, given that it has survived up to t; $\lambda = \frac{1}{\beta} > 0$
area under the density curve



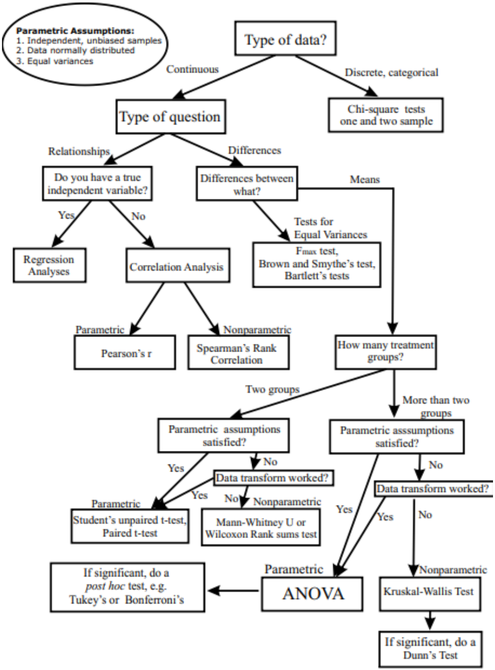
to the left of x $P\{X \leq x\} = 1 – e^{-\lambda x}$
to the right of x $P\{X > x\} = e^{-\lambda x}$
$P\{x1 < X \leq x2\} = e^{-\lambda x1} - e^{-\lambda x2}$

relation btw various dist

Relation between Bernoulli and Binomial Distribution 1. Bernoulli Distribution is a special case of Binomial Distribution with a single trial.
2. There are only two possible outcomes of a Bernoulli and Binomial distribution, namely success and failure.
3. Both Bernoulli and Binomial Distributions have independent trails.
Relation between Poisson and Binomial Distribution Poisson Distribution is a limiting case of binomial distribution under the following conditions:
The number of trials is indefinitely large or $\lim_{x \to \infty}$. The probability of success for each trial is same and indefinitely small or $\lim_{x \to 0}$. $np = \lambda$, is finite.

Poisson how many calls do you get in a day
The number of emergency calls recorded at a hospital in a day.
The number of thefts reported in an area on a day.
The number of customers arriving at a salon in an hour.
The number of suicides reported in a particular city.
The number of printing errors at each page of the book

exponential What about the interval of time between the calls
Length of time between metro arrivals,
Length of time between arrivals at a gas station
The life of an Air Conditioner
Exponential distribution is widely used for survival analysis.



1. $H_0 : \mu = 100$ and $H_1 : \mu \neq 100$
2. Rejection region is too far away from 100
3. if H_0 is true, how extreme is our sample?
4. Measure of extremeness, $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
5. Higher the z value, more likely to reject H_0

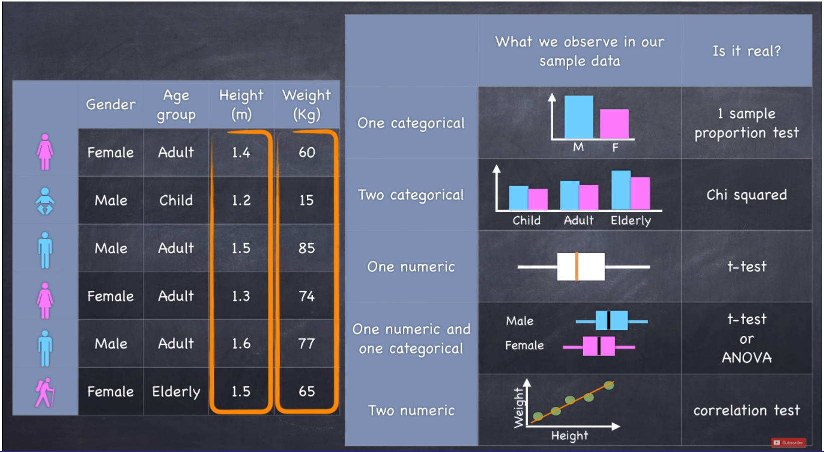
Confusion matrix

Type I and Type II Errors Type I error: Reject H_0 when H_0 is true
Prob (Type I Error) $=$ level of significance $= \alpha$ which is generally 5% Type II error: Not reject H_0 when H_0 is false
Prob (Type I Error) $=$ level of significance $= \beta = 1-$

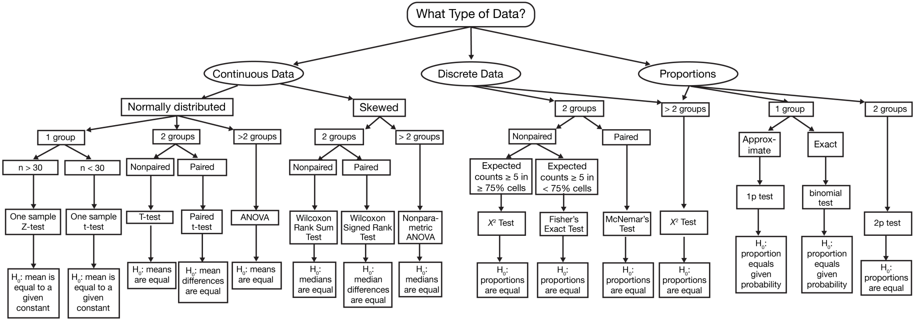
		CONDITION determined by "Gold Standard"			
	TOTAL POPULATION	CONDITION POS	CONDITION NEG	PREVALENCE $\frac{\text{CONDITION POS}}{\text{TOTAL POPULATION}}$	
TEST OUT-COME	TEST POS	True Pos TP	*Type I Error* False Pos FP	*Precision* Pos Predictive Value PPV = $\frac{\text{TP}}{\text{TEST P}}$	False Discovery Rate FDR = $\frac{\text{FP}}{\text{TEST P}}$
	TEST NEG	*Type II Error* False Neg FN	True Neg TN	False Omission Rate FOR = $\frac{\text{FN}}{\text{TEST N}}$	Neg Predictive Value NPV = $\frac{\text{TN}}{\text{TEST N}}$
	ACCURACY ACC ACC = $\frac{\text{TP + TN}}{\text{TOT POP}}$	*Sensitivity (SN), Recall* Total Pos Rate TPR TPR = $\frac{\text{TP}}{\text{CONDITION POS}}$	*Fall-Out* False Pos Rate FPR FPR = $\frac{\text{FP}}{\text{CONDITION NEG}}$	Pos Likelihood Ratio LR + LR + = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio DOR DOR = $\frac{\text{LR +}}{\text{LR -}}$
		Miss Rate False Neg Rate FNR FNR = $\frac{\text{FN}}{\text{CONDITION POS}}$	*Specificity (SPC)* True Neg Rate TNR TNR = $\frac{\text{TN}}{\text{CONDITION NEG}}$	Neg Likelihood Ratio LR - LR - = $\frac{\text{TNR}}{\text{FNR}}$	

Type I and Type II Errors Type I error: Reject H_0 when H_0 is true
Prob (Type I Error) $=$ level of significance $= \alpha$ which is generally 5% Type II error: Not reject H_0 when H_0 is false
Prob (Type I Error) $=$ level of significance $= \beta = 1-$

test statistics



Flow chart: which test statistic should you use?



When can AB test fail

1. in the case of a **referral program**, The referrer and Referee could be split across test and control groups causing spillover on the control or variant group
2. Novelty effects: Prompts and CTA tend to exhibit novelty effects, if not measuring their performance over the long term using a holdout a wrong attribution and/or customer fatigue can happen.
3. What-if scenarios: If you are looking to understand the impact of **not having launched a product**, for instance a subscription offering on a website. A/B test wouldn't be the right fit.

Class Imbalance

Univariate and Multivariate Analysis

Regression, Classification, Clustering

Regression

1. Linear
2. KNN
3. SVM
4. Random Forest

Classification

1. Logistic
2. KNN
3. SVM Classifier
4. Random Forest

Clustering

1. K-Means
2. Hierarchical
3. DBSCAN
4. HDBSCAN

Regression analysis is a statistical technique to assess the relationship between an predictor variable and one or more response factors.

Outcome Variable	**GLM Family**	**Link**	**Mean to Variance**
Continuous, unbounded	Normal or Standard Gaussian	Identity	
Continuous, non-negative	Gamma or inverse Gamma		
Discrete/ counts/ rate	Poisson Quassi-poisson or negative binomial	Log	Identity If not Identity
Count	Gamma		Over dispersion
Counts with multiple zero	Zero inflated poisson may be checked for fitting		
Binary	Binomial or Logistic regression		
Nominal	Multinomial regression		

Regression Model Selection Criteria

Three methods to classifier

1. model a classification rule - knn, decision tree, perceptron, svm
2. model the probability of class membership given input data - perceptron with cross-entropy cost
3. make a probabilistic model of data within each class - naive bayes 1 & 2 are discriminative classifications 3 is generative classification 2 & 3 probabilistic classification

content...

blocs

title of the bloc

bloc text

title of the bloc

bloc text

title of the bloc

bloc text

MACHINE LEARNING
chi square
big O notation
book - kevin murphy
Precision Recall tradeoff How to choose the method of predictive modelling. algorithms Bayesian Modelling (Topic Modelling), NLP, Bayesian Nonparametric Techniques, Social Network Analysis, Sentiment Analysis - https://www.springboard.com/blog/machine-learning-interview-questions/ -
https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms
https://blog.udacity.com/2016/04/5-skills-you-need-to-become-a-machine-learning-engineer.html
https://towardsdatascience.com/how-to-build-a-data-science-portfolio-5f566517c79c

Thank You!