Topics in Statistical Data Analysis: Revealing Facts From Data

Гэты сайт прапануе інфармацыю аб статыстычнага аналізу дадзеных. Ён апісвае аналіз часовых шэрагаў, папулярныя дыстрыбутывы, і іншыя тэмы. Ён разглядае выкарыстанне кампутараў у статыстычным аналізе дадзеных. У ім таксама пералічаныя кнігі, звязаныя і спасылкі на адпаведныя вэб-сайты.

Прафесар Хасейн Аршам

МЕНЮ

Companion сайтаў:

Для пошуку сайта, паспрабуйце E дит | F інд на старонцы [Ctrl+ F]. Калі ласка, увядзіце слова або фразу ў дыялогавым акне, напрыклад "Параметр " або "верагоднасць " Калі першае з'яўленне слова/фразу не тое, што вы шукаеце, паспрабуйце Eінд Далей.

Тэмы ў статыстычны аналіз дадзеных

Биномиального, паліномны, гипергеометрические, геаметрычныя, Pascal, адмоўны Значэнне, т размеркавання. Агульныя дыскрэтныя функцыі верагоднасці P-значэння для папулярных дыстрыбутываў

Якое ўплыў памеру?
Што такое закон Бенфорда? А як наконт закона Ципфа?
зрушэння метады скарачэння
Плошча пад стандартнай нармальнай крывой
нумар класа Interval ў Гістаграмы
Структурныя ўраўненні Мадэляванне
эканометрики і мадэляў часавых шэрагаў
трилинейной каардынат трохкутнік
ўнутраных і Inter-ацэншчык надзейнасці
Калі выкарыстоўваць непараметрические тэхнікі?
аналізу няпоўных дадзеных
Узаемадзеянне У дысперсійнай аналізу і Рэгрэсійная аналізу
кантрольных карт, а CUSUM
Six-Sigma якасці
паўтаральнасці і узнаўляльнасці
статыстычны інструмент, Grab адбору пробаў, і пасіўныя метады адбору пробаў
Адлегласць выбаркі
Байеса і эмпірычныя метады Байеса
і Маркоўскіх Памяць Тэорыя
верагоднасці Метады
дакладнасці, дакладнасці, надзейнасці і якасці
Функцыя ўплыву і яе дадатку
Што такое недакладнай верагоднасці?
Што такое мета-аналіз?
прамысловага мадэлявання дадзеных
прагнозу інтэрвал
Усталёўка дадзеных на ламанай
Як вызначыць, калі дзве лініі рэгрэсіі раўналежныя?
абмежаванае Рэгрэсійная мадэлі
Полупараметрические і не-параметрычнага мадэлявання
умеранасць і пасрэдніцтва
дискриминантного і класіфікацыя
індэкс падабенства ў класіфікацыі
абагульненых лінейных і лагістычных мадэляў
выжывання Аналіз
асацыяцыі паміж намінальнымі зменнымі
карэляцыі Спирмена і таў-прыкладанняў Кендал
паўторным вымярэннямі і падоўжныя дадзеныя
аналізу прасторавых дадзеных
Межы аналізу
Геостатистика мадэлявання
інтэлектуальнага аналізу дадзеных і выяўленне ведаў
нейронавых сетак Прыкладанні
тэорыі інфармацыі
захворвання і распаўсюджанасць
Выбар праграмнага забеспячэння
бокс-Кокса магутнасць пераўтварэнні
некалькіх параўнальных тэстах
Antedependent мадэлявання паўторных вымярэнняў
Спліт палове аналіз
паслядоўнага выбарачнага
лакальнага ўздзеяння
вариограмм Аналіз
крэдытнага скоринга: Спажывецкі крэдыт Ацэнка
кампанентаў працэнтныя стаўкі
частковых найменшых квадратаў
крывая росту Мадэляванне
насычаныя мадэлі і насычаныя Уваход Верагоднасць
Распазнанне вобразаў і класіфікацыя
Што такое биостатистики?
доказная Статыстыка
Статыстычныя судовай прыкладанняў
Што такое сістэматычны агляд?
Што такое Black-Sholes мадэль?
Што такое дрэва класіфікацыі?
Што такое рэгрэсія дрэва?
кластэрнага аналізу карэлявалі зменных
двайнога ахопу метады
Tchebysheff няроўнасці і яго паляпшэння
Фрэша мяжы для залежных выпадковых велічынь
Статыстычны аналіз дадзеных у галіне крымінальнага правасуддзя
Што такое інтэлектуальная лікавых разлікаў?
Software Engineering па кіраванні праектамі
Хі-квадрат Аналіз катэгарыяльны згрупаваных дадзеных
Каппа Коэна: Меры ўзгодненасці дадзеных
мадэлявання залежнасці катэгарыяльных дадзеных
Демінг Paradigm
Надзейнасць і рамонту сістэмы
вылічэнні стандартных вынікі
функцыі якасці Deployment (QFD)
падзей гісторыі аналіз

Увядзенне

Распрацоўкі ў галіне статыстычнага аналізу дадзеных, часта паралельна або наступных дасягненняў у іншых абласцях, да якіх статыстычныя метады плённа ўжываецца. Паколькі практыкі статыстычнага аналізу часта вырашэння канкрэтных прыкладных задач, рашэнне, метады развіцця, такім чынам, абумоўлена пошуку лепшага прыняцця рашэнняў ва ўмовах нявызначанасці.

Працэс прыняцця рашэнняў ва ўмовах нявызначанасці ў значнай ступені грунтуецца на прымяненні статыстычнага аналізу дадзеных для імавернасны ацэнкі рызыкі ваша рашэнне. Кіраўнікі павінны разумець змены па двух асноўных прычынах. Па-першае, каб яны маглі весці за сабой іншых ўжываць статыстычнага мыслення ў паўсядзённым дзейнасці і, па-другое, прымяніць паняцце з мэтай пастаяннага ўдасканалення. Гэты курс прадаставіць Вам практычны вопыт для садзейнічання выкарыстанню статыстычных метадаў мыслення і ўжываць іх, каб кампетэнтныя рашэнні, калі ёсць змяненні ў бізнес-дадзеных. Таму, вядома, у статыстычным мысленні праз дадзеныя-арыентаванага падыходу.

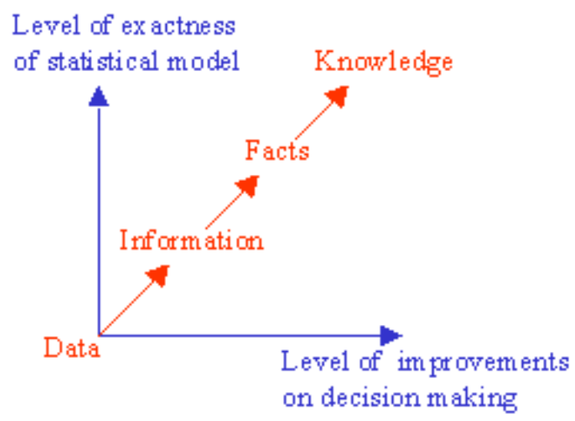
Статыстычныя мадэлі ў цяперашні час выкарыстоўваюцца ў розных галінах бізнесу і навукі. Тым не менш, тэрміналогія адрозніваецца ад поля да поля. Напрыклад, ўстаноўка мадэлі дадзеных, называецца каліброўка, гісторыі адпаведнасці, і засваенне дадзеных, з'яўляюцца сінонімамі з параметрам адзнакі.

Ваша арганізацыя базы даных змяшчае велізарную колькасць інфармацыі, аднак рашэнне членаў тэхналагічнай групы націсніце частка яе. Супрацоўнікі марнуюць час чысткі розных крыніц для базы дадзеных. Асоб, якія прымаюць рашэнні будуць расчараваныя, таму што яны не могуць атрымаць крытычна важных бізнэс-дадзеных, калі менавіта яны маюць патрэбу ў гэтым. Такім чынам, занадта многія рашэнні прымаюцца на аснове здагадак, а не факты. Шматлікія магчымасці і прапусціў, калі яны нават не заўважыў наогул.

Веданне таго, што мы добра ведаем. Інфармацыя з'яўляецца перадача ведаў. У кожным абмену ведамі, ёсць адпраўніком і атрымальнікам. Адпраўнік зрабіць агульнае, што з'яўляецца прыватным, не паведаміўшы, зносін. Інфармацыя можа быць класіфікавана як **відавочных і няяўных** формах. Дакладная інфармацыя можа быць растлумачана ў структураванай форме, у той час як маўклівае інфармацыі непаслядоўным і недакладным растлумачыць. Ведайце, што дадзеныя з'яўляюцца толькі сырая інфармацыя, а не веды самі па сабе.

Дадзеныя, як вядома, сырой інфармацыі, а не веды самі па сабе. Паслядоўнасць з дадзеных веды: **ад дадзеных да інфармацыі, ад інфармацыі факты, і, нарэшце, ад фактаў да ведаў**. Дадзеныя становяцца інфармацыяй, калі яна становіцца стаўленне да вашага рашэння праблемы. Інфармацыя становіцца фактам, калі дадзеныя могуць падтрымаць яго.Факты, што дадзеныя паказваюць. Аднак вырашальным інструментальныя (напрыклад, прыкладная) веды выяўляецца разам з некаторымі статыстычнымі ступень даверу.

Факт становіцца веданнем, калі яно выкарыстоўваецца ў паспяховым завяршэнні працэсу прыняцця рашэнняў. Калі ў вас ёсць велізарная колькасць фактаў, як інтэграваць веды, то ваш розум будзе звышчалавечае ў тым жа сэнсе, што чалавецтва з пісьмовага гэта звышчалавечыя ў параўнанні з чалавецтвам перад запісам. На наступным малюнку паказаны працэс статыстычнага мыслення, заснаванага на дадзеных пры пабудове статыстычных мадэляў для прыняцця рашэнняў ва ўмовах нявызначанасці.



Малюнку вышэй паказана, што як дакладнасць статыстычных мадэляў павялічваецца, узровень паляпшэнняў у працэс прыняцця рашэнняў ўзрастае. Вось чаму мы павінны статыстычнага аналізу дадзеных. Статыстычны аналіз дадзеных паўстала з неабходнасці паставіць веды на сістэматычнай доказнай базы. Гэта патрабуе вывучэння законаў верагоднасці, распрацоўка мер дадзеных уласцівасцяў і адносін, і гэтак далей.

Статыстычныя высновы накіравана на вызначэнні таго, любы статыстычнай значнасці можа быць далучаны, што вынікі пасля таго, як з-за ўліку любое выпадковае змяненне ў якасці крыніцы памылкі. Інтэлектуальны і крытычныя высновы не могуць быць тыя, хто не разумее мэты, умовы і прымяненне розных метадаў для ацэнкі значнасці.

Улічваючы нявызначанасць асяроддзя, верагоднасць таго, што "правільныя рашэнні" прымаюцца павялічваецца пры наяўнасці "дакладнай інфармацыі". Верагоднасць таго, што "добрая інфармацыя" можна азнаёміцца з павелічэннем ўзроўню структуравання працэсу кіравання ведамі. Вышэй паказчык таксама ілюструе той факт, што ў дакладнасці статыстычных мадэляў павялічваецца, узровень паляпшэнняў у працэс прыняцця рашэнняў ўзрастае.

Веды больш ведаючы, нешта тэхнічнае. мудрасць веды патрэбаў. Мудрасць улады пакласці наш час і нашы веды аб правільным выкарыстанні. Мудрасць прыходзіць з узростам і вопытам. Мудрасць ёсць дакладнае ўжыванне дакладных ведаў і іх ключавым кампанентам з'яўляецца ведаючы межы сваіх ведаў. Мудрасць пра ведаючы, як нешта тэхнічнае лепш за ўсё можа быць выкарыстаны для задавальнення патрэбнасцяў рашэнне. Мудрасць, напрыклад, стварае статыстычнае праграмнае забеспячэнне, што з'яўляецца карысным, а не тэхнічна бліскучым. Напрыклад, да гэтага часу вэб ўвайшоў у масавую свядомасць, назіральнікі адзначаюць, што яна ставіць інфармацыя ў вас пад рукой, але імкнецца захаваць мудрасць па-за дасяжнасцю.

Амаль усе спецыялісты павінны статыстычнага інструментара. Статыстычныя навыкі дазволяць Вам пісьменна збіраць, аналізаваць і інтэрпрэтаваць дадзеныя, якія датычацца іх вырашэння. Статыстычныя паняцці дазваляюць нам вырашаць задачы ў разнастайных сітуацыях. Статыстычнае мысленне дазваляе дадаваць рэчывы ў сваіх рашэннях.

З'яўленне кампутарных праграм, JavaScript, аплеты, аплеты Статыстычныя дэманстрацыі, і на сайце вылічэнняў найбольш важных падзей у працэсе выкладання і навучання ў канцэпцыі мадэлі на аснове статыстычных рашэнняў курсы. Гэтыя прылады дазваляюць будаваць лікавыя прыклады, каб

зразумець канцэпцыі, і знайсці іх значнасць для сябе.

Мы будзем прымяняць асноўныя паняцці і метады статыстыкі вы ўжо даведаліся з папярэдняга курсу статыстыкі рэальных праблем свету. Курс спецыяльна распрацаваны для задавальнення вашых патрэбаў ў статыстычных дадзеных бізнэс-аналізу з дапамогай шырока даступных *камерцыйных* статыстычных пакетаў кампутара, такія як SAS і SPSS.Робячы гэта, вы непазбежна апынецеся задаваць пытанні аб дадзеных, і прапанаваны метад, і вы будзеце мець сродкі ў Вашым распараджэнні, каб урэгуляваць гэтыя пытанні, да ўласнага задавальнення. Адпаведна, усё прыкладанні, праблемы запазычаных з бізнесу і эканомікі. Да канца гэтага курсу Вы будзеце ўмець думаць статыстычна пры выкананні любога аналізу дадзеных.

Ёсць дзве агульныя погляды выкладання/вывучэння статыстыкі: Вялікага і Малога статыстыкі. Больш статыстыкі ўсе звязаныя з навучаннем па дадзеных, ад першага планавання і збору, на апошняй прэзентацыі або справаздачы. Малы статыстыкі цела статыстычнай метадалогіі. Гэта *Вялікі курс статыстыкі*.

Існуюць два асноўных выгляду "статыстыка" курсы. Рэальны выгляд паказвае, як разабрацца ў дадзеных. Гэтыя курсы будуць ўключаць у сябе ўсе апошнія падзеі і ўсе падзяляем глыбокую павагу да дадзеных і праўды. Імітацыя выгляд ўключае ў сябе падключэнне нумары ў статыстыку формул. Акцэнт робіцца на тым, арыфметычныя правільна. Гэтыя курсы, як правіла, не зацікаўлены ў дадзеных або праўда, і праблемы, як правіла, арыфметычныя практыкаванні. Калі пэўныя здагадкі, неабходныя для апраўдання працэдуры, яны будуць проста сказаць вам "ўзяць на сябе... нармальна размеркаваныя", - незалежна ад таго, як малаверагодна, што магло б быць. Падобна, вы пакутуеце ад перадазіроўкі апошняга. Гэты курс будзе вывесці радасць статыстыкі ў вас.

Статыстыка ёсць навука дапамагчы вам у прыняцці рашэнняў ва ўмовах нявызначанасці (на аснове лікавых і вымерна маштабах). Працэс прыняцця рашэнняў павінен быць заснаваны на дадзеных, ні на асабістае меркаванне, ні па веры.

Гэта ўжо прызнаным фактам, што "Статыстычнае мысленне калі-небудзь будзе па меры неабходнасці для эфектыўнага грамадзянства, як уменне чытаць і пісаць". Такім чынам, давайце будзем наперадзе нашага часу.

Папулярных дыстрыбутываў і іх Тыповыя вобласці ўжывання

Биномиальный

Ужыванне: Дае верагоднасць менавіта поспехі ў п незалежных выпрабаваннях, калі верагоднасць поспеху р у адным выпрабаванні з'яўляецца канстантай. Часта выкарыстоўваецца для кантролю якасці, надзейнасці, выбарачнае абследаванне, а таксама іншых вытворчых задач.

Прыклад: Якая верагоднасць таго, у 7 і больш "галавы" на 10 кідкоў справядлівай манеты?

Каментарыі: часам можна аппроксимировать нармальным або размеркаваннем Пуасона.

Паліномны

Ужыванне: Дае верагоднасць роўна п я выніках падзеі я, да = 1, 2,..., А ў п незалежных выпрабаваннях, калі верагоднасць р я падзеі я ў *1* судовы працэс з'яўляецца пастаянным.Часта выкарыстоўваецца ў галіне кантролю якасці і іншых вытворчых задач.

Прыклад: Чатыры кампаніі таргі для кожнага з трох кантрактаў з названай верагоднасці поспеху. Якая верагоднасць таго, што адна кампанія атрымае ўсё заказы?

Каментары: абагульненне биномиального размеркавання руды больш за 2 вынікаў.

Гипергеометрические

Ужыванне: Дае верагоднасць выбару менавіта х добры адзінак у выбарцы з п элементаў з сукупнасці адзінак N, калі ёсць да дрэннай адзінак насельніцтва. Выкарыстоўваецца ў галіне кантролю якасці і звязаных з імі прыкладанняў.

Прыклад: Улічваючы, шмат добрага на 21 адзінак і чатыры няспраўны. Якая верагоднасць таго, што выбарка з пяці дасць не больш адной бракаванай?

Каментарыі: Можа быць аппроксимирована биномиального размеркавання, калі я мала звязаныя з N.

Геаметрычны

Ужыванне: Дае верагоднасць якія патрабуюць дакладнасці х биномиальных выпрабаванняў, перш чым першы поспех дасягнуты. Выкарыстоўваецца ў кантролі якасці, надзейнасці і іншых прамысловых аб'ектах.

Прыклад: Вызначэнне верагоднасці патрабуецца роўна пяць тэстаў стрэльбаў перад першым поспехам дасягнута.

Паскаль

Ужыванне: Дае верагоднасць сапраўды х няўдач папярэдніх н поспехам.

Прыклад: Якая верагоднасць таго, што трэці поспех мае месца на 10-й суд?

Адмоўнае биномиальное

Ужыванне: Дае верагоднасць аналагічныя размеркавання Пуасона, калі падзеі не адбываюцца з пастаяннай хуткасцю, а хуткасць з'яўлення з'яўляецца

выпадковай велічынёй, якая варта гама-размеркаванне.

Прыклад: Размеркаванне колькасці паражнін для групы стаматалагічных пацыентаў.

Каментары: Абагульненне размеркаванне Паскаля, калі з не з'яўляецца цэлым лікам. Многія аўтары не робяць адрозненні паміж Паскалем і адмоўнага биномиального размеркавання.

Пуасона

Ужыванне: Дае верагоднасць сапраўды х незалежных уваходжанняў на працягу пэўнага перыяду часу, калі падзеі адбываюцца незалежна адзін ад аднаго і з пастаяннай хуткасцю.Таксама можа прадстаўляць ліку з'яўленняў на пастаяннай плошчаў або аб'ёмаў. Часта выкарыстоўваецца для кантролю якасці, надзейнасці, тэорыі масавага абслугоўвання, і гэтак далей.

Прыклад: Выкарыстоўваецца для прадстаўлення размеркавання колькасці дэфектаў у частцы матэрыялаў, якія прыбылі кліентаў, страхавыя выплаты, якія ўваходзяць тэлефонныя званкі, альфа-часціц, выпусканых, і гэтак далей.

Каментарыі: Часта выкарыстоўваецца як набліжэнне биномиального размеркавання.

Нармальны

Ужыванне: асноўнае размеркаванне статыстыкі. Шматлікія прыкладанні звязаны з цэнтральнай лімітавай тэарэмай (у сярэднім значэння назіранняў п. набліжаецца да нармальнага размеркаванні, незалежна ад формы арыгінальнага дыстрыбутыва пры дастаткова агульных умовах). Такім чынам, падыходнай мадэллю для многіх, але не ўсе фізічныя з'явы.

Прыклад: Размеркаванне фізічных вымярэнняў на жывых арганізмах, тэсты на інтэлект, памеры вырабаў, сярэдняя тэмпература, і гэтак далей.

Каментарыі: Шматлікія метады статыстычнага аналізу мяркуюць нармальнае размеркаванне.

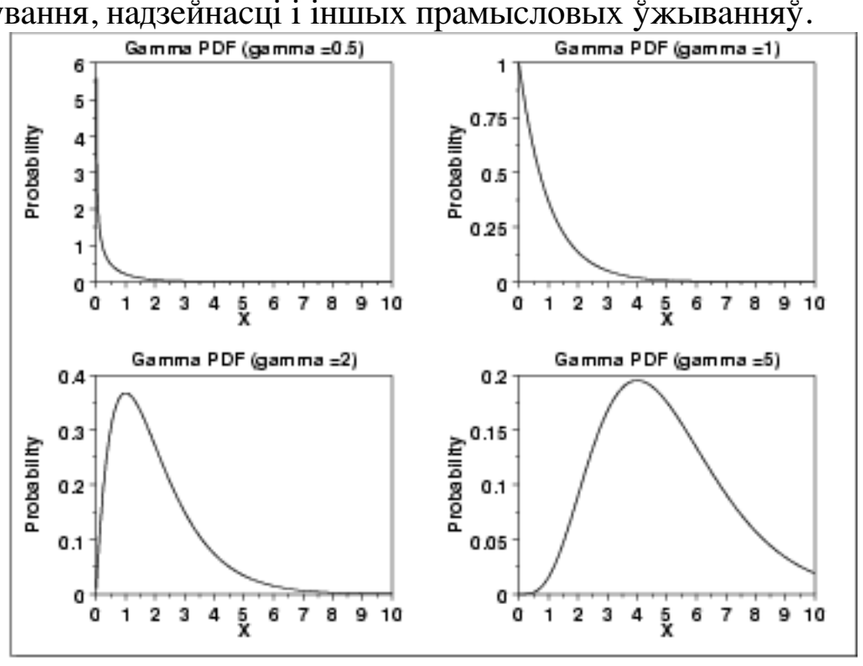
Так званыя абагульненыя размеркаванне Гаўса мае наступныя PDF:

A.exp [B | x | N], дзе A, B, N-пастаянныя. Пры п = 1 і 2, Лапласа і размеркаванне Гаўса адпаведна. Гэта размеркаванне набліжаецца досыць добрыя дадзеныя ў некаторых кадавання малюнкаў, прыкладанняў.

Slash размеркавання з'яўляецца размеркаванне стаўленне нармальнай выпадковай велічыні з раўнамерным незалежнай выпадковай велічыні, гл Т. Хатчынсан, *бесперапынная двухмерных размеркавання*, Rumsby праф. Публікацыі 1990 года.

Гама

Ужыванне: асноўнае размеркаванне статыстыкі для зменных, абмежаваная з аднаго боку - напрыклад, х больш або роўная нулю. Дае размеркаванне часу, неабходнага для дакладна да незалежных падзеі адбываюцца, мяркуючы, што падзеі адбываюцца з пастаяннай хуткасцю. Часта выкарыстоўваецца ў тэорыі масавага абслугоўвання, надзейнасці і іншых прамысловых ўжыванняў.



Прыклад: Размеркаванне часу паміж паўторнай каліброўкі інструментаў, якую неабходна паўторна пасля каліброўкі да выкарыстоўвае, час паміж інвентарызацыі папаўненне, напрацоўка на адмову для сістэмы з рэжыму чакання кампанентаў.

Каментары: Эрланген, экспанентнае, і хі-квадрат размеркавання асаблівых выпадках. Дирихле з'яўляецца шматмерным пашырэннем Бэта размеркавання.

Размеркаванне прадуктаў незалежных аднолькава раўнамернай (0, 1) выпадковых? Як і многія іншыя праблемы з прадуктамі, гэта становіцца праблемай, калі знаёмыя ператварыліся ў задачы аб сумах. Калі X з'яўляецца аднастайным (для прастаты пазначэнняў робяць U (0,1)), Y = часопіса (X) размеркаваны экспанентна, так што часопіс творы X1, X2,... X ўяўляе сабой суму Y1, Y2,... Уп які мае гама-выпраменьвання (маштабуецца хі-квадрат) размеркаванне. Такім чынам, гама шчыльнасці з параметрам формы п, маштаб 1.

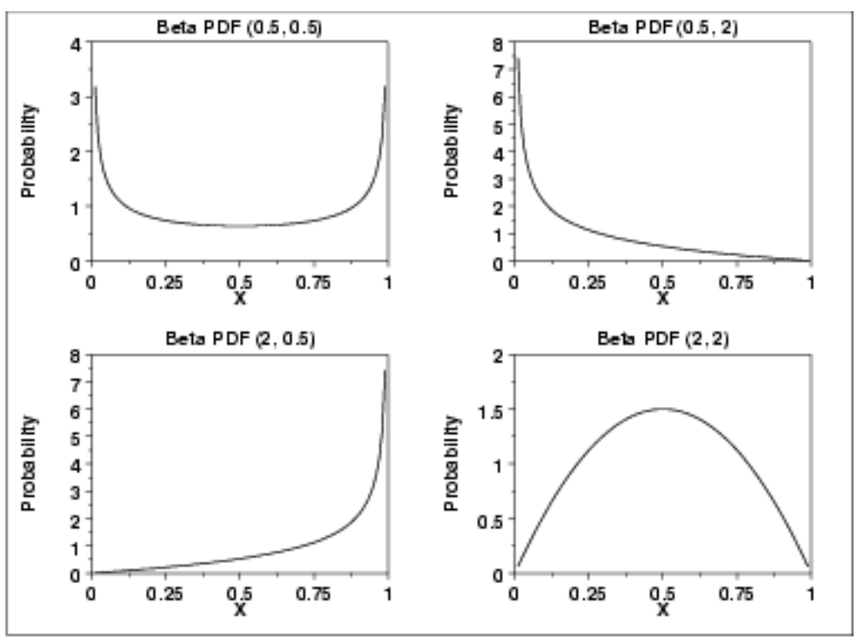
Экспанентны

Ужыванне: дае размеркаванне часу паміж незалежнымі падзеямі, якія адбываюцца з пастаяннай хуткасцю. Гэта раўнасільна таму, размеркаванне верагоднасцяў жыцця, мяркуючы, пастаянны умоўны адмова (або небяспекі) стаўцы. Такім чынам, дастасавальныя ў многіх, але не ўсе сітуацыі надзейнасць.

Прыклад: Размеркаванне часу паміж прыбыццём часціц на лічыльнік. Акрамя таго, жыццё размеркавання комплекс неизбыточных сістэм, а тэрмін эксплуатацыі некаторых кампанентаў - у прыватнасці, калі яны падвяргаюцца пачатковай выгарання, а таксама прафілактыкі ліквідуе частцы да зносу.

Каментары: Асаблівы выпадак, так Вейбулла і гама-размеркавання.

Бэта



Ужыванне: асноўнае размеркаванне статыстыкі для зменных, абмежаваных з абодвух бакоў - напрыклад, паміж х аб і 1. Карысна для тэарэтычных і прыкладных задач у розных галінах.

Прыклад: Размеркаванне долі насельніцтва, размешчаных паміж мінімальным і максімальным значэннем у узоры, размеркаванне штодзённых адсоткаў прыбытковасці ў вытворчым працэсе, апісанне які прайшоў раз завяршэння задачы (PERT).

Каментары: Уніформа, правая трохкутная, і парабалічнага размеркавання асаблівых выпадках. Каб стварыць бэта стварэння двух выпадковых велічынь з гамы, г 1, г 2. Стаўленне г 1 /(г 1+ г 2) распаўсюджваецца як бэта-размеркавання. Бэта-размеркаванне можа таксама разглядацца як размеркаванне X1 дадзеных (X1+ X2), калі X1 і X2 незалежных гама выпадковых велічынь.

Існуе таксама сувязь паміж Beta і нармальныя размеркавання. Традыцыйны разлік, што пры PERT Beta з вышэйшай каштоўнасцю, як бы нізкая, як і, хутчэй за ўсё, як м, што эквівалентна нармальнае размеркаванне мае сярэднюю і рэжым (+ 4M+ б)/6 і стандартнае адхіленне (бы - а)/6.

Гл. раздзел 4.2, *увядзенне ў верагоднасці* Дж. Лоры Снелл (Нью-Ёрк, Random House, 1987) для сувязі паміж бэта-і F размеркавання (з тым перавагай, што табліцы лёгка знайсці).

Аднастайны

Ужыванне: дае верагоднасць таго, што назіранне будзе адбывацца на працягу пэўнага інтэрвалу, калі верагоднасць з'яўлення ў гэтым інтэрвале прама прапарцыйная працягласці інтэрвалу.

Прыклад: Выкарыстоўваецца для генерацыі выпадковых шануюць.

Каментары: Асаблівы выпадак бэта-размеркавання.

Шчыльнасць сярэдняе геаметрычнае з п незалежных форме (0,1):

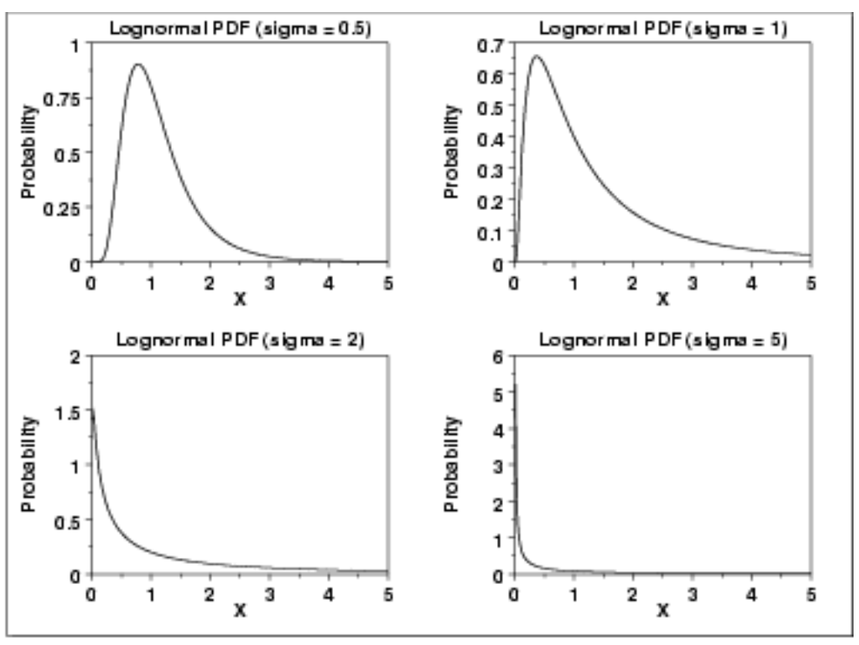
P (X = x) = п (п-1) (Уваход [1/х л]) (п-1) /(п-1)!.

г L = [U L - (1-U) L]/L, як кажуць, сіметрычны Тьюки л -размеркавання.

Уваход для нармальнай

Ужыванне: Дазволы ўяўленне выпадковай велічыні, лагарыфм якога варта нармальнаму размеркаванні. Мадэль працэсу, якія вынікаюць з шматлікіх дробных памылак мультыплікатыўны. Адпаведныя калі значэнне назіранай зменнай з'яўляецца выпадковай доля назіраліся раней значэння.

У выпадку, калі дадзеныя логнормально размеркаванай геаметрычнай актаў сярэдняе, як лепш дэскрыптар, чым сярэдняе. Чым бліжэй дадзеныя вынікаюць логнормального размеркавання, тым бліжэй сярэдняе геаметрычнае з'яўляецца сярэдняй, так як часопіс паўторнага выразы стварае сіметрычнае размеркаванне.



Прыклад:. Размеркаванне памераў ад паломкі працэсу размеркавання даходаў памеру спадчыны і банкаўскія дэпазіты; размеркавання розных

біялагічных з'яў, жыццё распаўсюджвання некаторых тыпаў транзістараў
адносіны дзвюх лог-нармальнае размеркаванне зменных лог-нармальнай.

Рэлея

Ужыванне: дае размеркаванне радыяльнай памылкі, калі памылкі ў двух узаемна перпендыкулярных восяў незалежныя і нармальна размеркаваныя вакол нуля з роўнымі дысперсіямі.

Прыклад: Бомба-прыцэльнай праблем, амплітуда шуму канверт, калі лінейны датчык выкарыстоўваецца.

Каментары: Асаблівы выпадак Вейбулла.

Коши

Ужыванне: дае размеркаванне адносіны двух незалежных стандартызаваных нармальных зменнымі.

Прыклад: Размеркаванне дачыненні да стандартызаваных чытанняў шум, размеркавання загару (х) пры х раўнамерна размеркавана.

Хі-квадрат

Крывая шчыльнасці верагоднасці размеркавання хі-квадрат з'яўляецца асіметрычнай крывой працягласцю больш станоўчы бок лініі і маюць працяглы права хвост. Форма крывой залежыць ад значэння ступеняў свабоды.

Вобласці ўжывання: Найбольш шырока прымяненне хі-квадрат размеркаванне, з'яўляюцца:

- Хі-квадрат для Асацыяцыі (непараметрический, таму можа быць выкарыстаны для намінальных дадзеных) праверкі статыстычнай значнасці шырока выкарыстоўваюцца двухмерныя таблічныя асацыяцыі аналізу. Як правіла, гіпотэза або няма двух розных папуляцый досыць розныя ў некаторых характэрных або аспект іх паводзінаў, заснаваных на двух выпадковых выбарак. Гэтая працэдура выпрабаванняў таксама вядомы як Pearson хі-квадрат.
- Хі-квадрат дабра згоды, тэст выкарыстоўваецца для праверкі, калі назіранае размеркаванне адпавядае якому-небудзь канкрэтнаму размеркаванні. Разлік гэтага крытэрыю згоды тэст параўнання дадзеных назіранняў з дадзенымі чакаць на аснове канкрэтнага размеркавання.

Вейбулла

Ужыванне: Агульны час да адмовы размеркаванні ў сувязі з шырокім разнастайнасцю небяспекі курс крывых і экстрэмальных значэнняў размеркавання мінімальнага значэння N ад размеркавання абмежаваных злева.

Размеркаванне Вейбулла часта выкарыстоўваецца для мадэлі "час да адмовы". Такім чынам, ён ужываецца ў актуарная навукі і інжынерныя працы.

Гэта таксама адпаведнае размеркаванне для апісання дадзеных, якія адпавядаюць рэзананснага паводзінаў, такія як змяненне з энергіяй перасеку ядзерных рэакцый або змяненне з хуткасцю паглынання выпраменьвання ў эфекце Мессбауэра.

Прыклад: Жыццё размеркавання для некаторых кандэнсатараў, падшыпнікі, рэле, і гэтак далей.

Каментары: Рэлея і экспанентнае размеркавання з'яўляюцца прыватнымі выпадкамі.

Экстрэмум

Ужыванне: Абмежаванне мадэль размеркавання максімальных або мінімальных значэнняў N выбіраецца з "экспанентнае тыпу" размеркавання, такія як нармальнае, гама, або экспанентным.

Прыклад: Размеркаванне трываласць некаторых матэрыялаў, кандэнсатара напружанне прабоя, парыў хуткасці сутыкнуліся самалёты, раз бактэрыі знікнення.

т размеркавання

Т размеркавання былі выяўленыя ў 1908 годзе Уільям Гос, які быў хімікам і статыстыка, якія працуюць у піваварнай кампаніі Guinness. Ён лічыў сябе студэнтам яшчэ вучыцца статыстыкі, так што гэта, як ён падпісваў свае дакументы, як псеўданімам "Студэнт". Ці, магчыма, ён выкарыстаў псеўданім ў сувязі з "камерцыйнай таямніцы" абмежаванняў Гінеса.

Звярніце ўвагу, што існуюць розныя дыстрыбутывы т, то клас размеркаванняў. Калі мы гаворым аб канкрэтным размеркаванні т, мы павінны паказаць ступеняў свабоды. Т крывых шчыльнасці сіметрычныя і званіца форму нармальнага размеркавання, і іх пік на 0. Тым не менш, роскід больш, чым у стандартнага нармалёвага размеркавання. Чым больш ступеняў свабоды, чым бліжэй Т-шчыльнасць нармальнай шчыльнасці.

Чаму кожная рэч кошце аднаго пені на долар?

Вось псіхалагічны адказ. У сувязі з вельмі абмежаванай магчымасцю апрацоўкі дадзеных мы, людзі спадзяюцца на катэгарызацыі (напрыклад, бачыць рэчы, як "чорнае або белае" патрабуе толькі двайковыя схемы кадавання, а не бачыць мноства адценняў шэрага). Наша сістэма злічэння мае асноўнай катэгорыі 100 (напрыклад, 100 пені, 200 пені, 300 пені) і ёсць афектыўных адказ, звязаных з гэтымі групамі - чым больш, тым лепш, калі вы атрымліваеце іх, больш дрэнна, калі вы даеце ім. Рэклама і цэны перавага гэтага абмежаванага апрацоўкі дадзеных на $ 2.99, $ 3.95, і г.д. Так што $ 2,99 нясе афектыўнай адказу звязана з 200-група капейкі. Сапраўды, калі вы спытаеце людзей у адказ на "як блізка адзін да аднаго" з'яўляюцца 271 і 283 у параўнанні з "як блізка адзін да аднаго", 291 і 303, былы лічацца бліжэй (ёсць шмат методык створаны, каб адгаварыць суб'ектаў проста адняць менш ад большага). Акрамя таго, шкода, праца акцыях, спартыўных спаборніцтвах, а таксама мноства іншых актывізуе спробы звязаць

вялікія якасныя адрозненні з тым, што часцяком нязначныя колькасныя адрозненні, напрыклад, металічнага золата ў алімпійскіх падзей басейн можа быць мілісекунд адрозненне ад не метал.

Яшчэ адна матывацыя: Псіхалагічна $ 9.99 можа выглядаць лепш, чым $ 10.00, але ёсць больш фундаментальныя прычыны таксама. Памочнік павінен даць вам перайсці ад вашага 10 даляраў, і павінен тэлефанаваць у продажы праз яго/яе грошы зарэгіструйцеся, каб атрымаць на 1 цэнт. Гэта прымушае здзелка, каб прайсці праз кнігі, вы атрымаеце квітанцыю, і памочнік не можа проста забіць яму $ 10/сама. Майце на ўвазе, што няма нічога, каб спыніць асабліва ненадзейным супрацоўнікам ўдаючыся ў працы з кішэню цэнтаў...

Там ёсць падатак з продажаў для гэтага. Для любой цаной (прынамсі ў ЗША), вы павінны будзеце заплаціць падатак з продажаў таксама. Так што вырашае праблему адкрыцця касавага апарата. Гэта, а таксама камеры відэаназірання;).

Там былі некаторыя даследаванні ў тэорыі маркетынгу на **паводзіны спажыўца** у прыватнасці цане. Па сутнасці, яны звязаны з пакупніком чаканнямі, заснаванымі на папярэднім вопыце. Крытычнае даследаванне выпадак у Вялікабрытаніі па кошце указаннем калготкі (панчохі) паказалі, што было ярка выяўленых пікаў попыту на пакупніка чаканых цэнавых з 59р, 79 пенсаў, 99р, £ 1,29 і гэтак далей. Попыт на прамежкавых кропках цана была значна ніжэй чаканай гэтых кропак на аналагічны тавар якасцю. У Вялікабрытаніі, напрыклад, цэны на віна звычайна ўсталёўваюцца ў ключавых кропках кошты. Віно рознічнага гандлю таксама пацвярджаюць, што продажы па розных цэнах (нават капейкі ці такія розныя) не прыводзяць да зусім розныя аб'ёмы продажаў.

Іншыя даследаванні паказалі, што наадварот, дзе зніжэнне коштаў паказаў зніжэнне аб'ёмаў продажаў, спажыўцоў, прыпісваючы якасці ў адпаведнасці з цаной. Тым не менш, ён не цалкам пратэставаны, каб вызначыць, аб'ём продажаў працягвае расці з цаной.

Іншыя падобныя даследаванні аказваецца на паводзіны спажыўцоў да зменаў цэн. Ключавым пытаннем тут з'яўляецца тое, што проста прыкметная розніца (JND), ніжэй якой спажыўцы не будуць дзейнічаць на павышэнне коштаў. Гэта мае практычнае прымяненне пры павелічэнні ставак збораў і таму падобнае. JND, як правіла, 5%, і гэта дае магчымасць для кансультантаў і г.д. для павышэння коштаў вышэй, да стаўкі не больш чым на JND без скаргі кліента. У якасці эмпірычнага эксперыменту, паспрабуйце паразарадкі кліентаў на 1, 2,.., 5, 6% і назіраць за рэакцыяй. Да 5%, як уяўляецца, не аказвае адмоўнага ўздзеяння.

З іншага боку, няма ніякага сэнсу ў прапанове платы скарачэнне складае менш за 5% кліентаў не прызнаюць канцэсіі вы зрабілі. Аналагічным чынам, у перыяды інфляцыі, росту коштаў павінны быць арганізаваны так, што чалавек ростам цэн знаходзіцца пад 5%, магчыма, за кошт павышэння коштаў на 4%, два разы на год, а не адзін ад 8% росту.

Кароткая гісторыя верагоднасцяў і матэматычнай статыстыкі

Арыгінальная ідэя "статыстыка" стаў збор інфармацыі аб і "дзяржава". Слова статыстыку дыскі непасрэдна не з класічных грэцкіх і лацінскіх каранёў, але ад італьянскага словастану.

Нараджэнне статыстыкі адбылося ў сярэдзіне 17 -га стагоддзя. Незнатного паходжання, па імені Джон Граунт, які быў родам з Лондана, пачаць разгляд штотыднёвік царква выдадзеных мясцовымі клерк прыход, што пералічаныя колькасць нараджэнняў, хрэсьбіны і смяротнасці ў кожным прыходзе. Гэтыя так званыя вэксалі Смяротнасць таксама ўказаны прычыны смерці. Граунт, які быў уладальнікам крамы арганізаваны гэтыя дадзеныя ў формах, якія мы называем апісальнай статыстыкі, якая была апублікаваная ў якасці*прыроднага і палітычныя назірання, зробленыя пры законапраектаў аб смяротнасці*. Неўзабаве пасля гэтага ён быў абраны членам Каралеўскага грамадства. Такім чынам, статыстыка павінна пераймаць шэраг паняццяў ад сацыялогіі, напрыклад, паняцце "насельніцтва". Было выказана меркаванне, што, паколькі статыстыку звычайна ўключае ў сябе вывучэнне чалавечага паводзінаў, яна не можа прэтэндаваць на дакладнасць фізічных навук.

Верагоднасць мае значна больш даўнюю гісторыю. верагоднасці паходзіць ад дзеяслова, каб даследаваць сэнс "даведацца", што не так ужо і лёгка даступныя і зразумелыя. Слова "доказ" мае тое ж паходжанне, што дае неабходную інфармацыю, каб зразумець тое, што сцвярджаў, каб быць праўдай.

Верагоднасць паўстаў з вывучэння азартных гульняў і ігральных ў шаснаццатым стагоддзі. Тэорыя верагоднасцяў быў профіль матэматыкі вывучаюцца Блез Паскаля і П'ера Ферма ў 17. Стагоддзі. Цяперашні час, у 21 -м стагоддзі, імавернасны мадэлі, якія выкарыстоўваюцца для кіравання патокам трафіку праз шашы сістэмай, тэлефонам абмену, або працэсар кампутара, знайсці генетычны склад асобных асоб або груп насельніцтва; кантроль якасці, страхаванне, інвестыцыйныя і іншыя сектары бізнесу і прамысловасці.

Новыя і пастаянна расце розных галінах чалавечай дзейнасці, выкарыстоўваюць статыстыку, аднак, здаецца, што гэта само поле застаецца незразумелым для шырокай публікі.Прафесар Брэдлі Эфрон выказаў гэты факт добра:

> На працягу 20 -га стагоддзя статыстычнага мыслення і метадалогіі сталі навуковай асновай для літаральна дзясяткі абласцей, у тым ліку адукацыя, сельская гаспадарка, эканоміка, біялогія і медыцына, і ўсё большы ўплыў у апошні час на дакладныя навукі, такія як астраномія, геалогія, фізіка. Іншымі словамі, мы выраслі з маленькай смутнай поля ў вялікі цёмны поле.

Дадатковая літаратура:

Daston Л., *класічнай верагоднасці ў эпоху Асветы*., Princeton University Press, 1988
У кнізе адзначаецца, што старажытныя мысляры Асветы не можа сутыкнуцца з нявызначанасцю. Механистический, дэтэрмінаваных машыне, быў Асветы погляд на свет.
Гиллис Д., *філасофскія тэорыі верагоднасцяў*, Routledge, 2000. Вокладкі класічных, лагічнага, суб'ектыўнага, частату і схільнасць думкі.
Узлом І., *Узнікненне тэорыі верагоднасцяў*, Cambridge University Press, London, 1975. Філасофскага даследавання ранніх уяўленняў аб верагоднасці, індукцыі і статыстычныя высновы.
Peters W., *для падліку что-то Статыстычныя прынцыпы і асобы*, Springer, Нью-Ёрк, 1987 год. Яна вучыць прынцыпам прыкладной эканамічнай і сацыяльнай статыстыкі ў гістарычным кантэксце. Рэкамендуемыя тэмы ўключаюць у сябе апытанні грамадскай думкі, вытворчы кантроль якасці, фактарнага аналіз, байесовских метадаў, ацэнка праграм, непараметрические і надзейных метадаў і разведачны аналізу дадзеных.
Портер Т. *Павышэнне статыстычнага мыслення*, 1820-1900, Princeton University Press, 1986 год. Аўтар сцвярджае, што статыстыка стала вядомая ў XX стагоддзі ў якасці матэматычнага апарату для аналізу эксперыментальных і назіральных дадзеных. Замацаваных у дзяржаўнай палітыцы ў якасці адзінай надзейнай асновай для меркаванні аб эфектыўнасці медыцынскіх працэдур або бяспекі хімічных рэчываў, а таксама прынятыя **бізнесу** для такіх ужыванняў, як кантроль якасці прамысловых, яна, відавочна, сярод прадуктаў навука, чыё ўплыў на грамадскую і прыватнае жыццё быў найбольш распаўсюджаным. Статыстычны аналіз таксама прыйшоў, каб убачыць ў шматлікіх навуковых дысцыплін з'яўляюцца абавязковымі для складання надзейных высноў з эмпірычных results.This новую вобласць матэматыкі выявілі гэтак шырокія вобласці прымянення.
Стиглер С., *Гісторыя статыстыкі: вымярэнне нявызначанасці Да 1900*, У. Chicago Press, 1990. Ён ахоплівае людзей, ідэй і падзей, якія ляжаць у аснове ўзнікнення і развіцця ранніх статыстыкі.
Tankard Дж. *Статыстычныя піянераў*, Schenkman Кнігі, Нью-Ёрк, 1984 год.
Гэтая праца дае падрабязную жыццё і часы тэарэтыкі, праца якіх працягвае фармаваць шмат сучаснай статыстыкі.

Розныя школы думкі ў галіне статыстыкі

Ёсць некалькі розных школ думкі ў галіне статыстыкі. Яны ўводзяцца паслядоўна ў часе па неабходнасці.

Працэс нараджэння новай школы думкі

Працэс распрацоўкі новай школы думкі ў любой вобласці заўсёды займаў натуральны шлях. Нараджэнне новай школы думкі ў статыстыцы не з'яўляецца выключэннем. Працэс нараджэння прыводзіцца ніжэй:

З улікам ужо устаноўленых школу, трэба працаваць у рамках пэўнай структуры.

Крызіс з'яўляецца, гэта значыць некаторыя неадпаведнасці ў рамках выніку яго уласным законах.

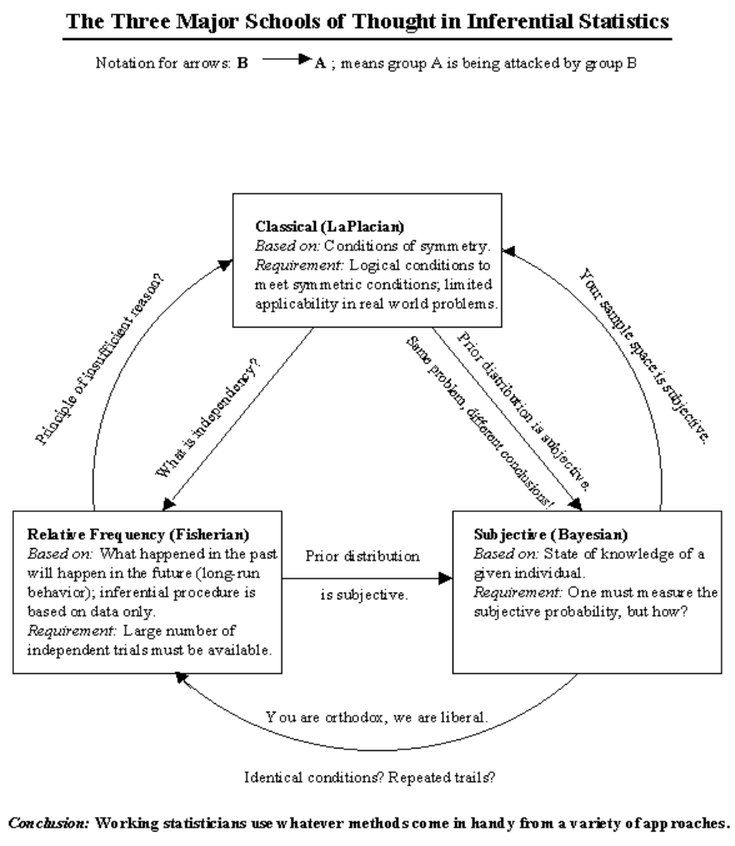
Адказ паводзін:

1. Нежаданне лічыцца з крызісам.
2. Паспрабуйце размясціць і растлумачыць крызіс у існуючых рамках.
3. Пераўтварэнне некаторых вядомых навукоўцаў прыцягвае паслядоўнікаў у новую школу.

Успрыманне крызісу ў статыстычным супольнасці выклікае патрабуе для "падмурак ўмацоўваюць". Пасля таго, як крызіс скончыцца, усё можа выглядаць інакш, і гісторыкі статыстыка можа прывесці падзеі, як у серыі крокаў "абапіраючыся на падмурак". Такім чынам, мы можам прачытаць гісторыю статыстыкі, гісторыі пірамиду стварыў пласт за пластом на трывалую аснову з цягам часу.

Іншыя школы думкі ўзнікаюць па пашырэнні і "змякчыць" існуючай тэорыі верагоднасцяў і статыстыкі. Некаторыя "змякчэння" падыходы выкарыстоўваюцца паняцці і метады, распрацаваныя ў тэорыі недакладных множстваў, тэорыя магчымасці і Дэмпстер-Шафер тэорыя.

На наступным малюнку паказаны тры асноўныя школы думкі, а менавіта: класічны (прыпісваецца Лапласа), адносная частата (прыпісваецца Fisher) і байесовский (прыпісваецца Savage). Стрэлкі на гэтым малюнку з'яўляюцца аднымі з асноўных прэтэнзій у мэты, частотнай і суб'ектыўнай школы. У якой школе вы належыце? Прачытайце выснову на гэтым малюнку.



Які тып статыстык Вы? **Націсніце на малюнак, каб павялічыць**

Дадатковая літаратура : Платон, Ян фон, *стварэнне сучаснай верагоднасці*, Cambridge University Press, 1994. Гэтая кніга ўяўляе сабой гістарычную пункт гледжання на субъективистской і объективистской верагоднасць школы думкі. Прэс С., Дж. Tanur, *суб'ектыўнасць навукоўцаў і байесовский падыход*, М., 2001. Параўнанне і проціпастаўленне рэальнасці суб'ектыўнасці ў працы вялікіх вучоных у гісторыі і сучаснай байесовский падыход да статыстычнаму аналізу. Weatherson Б., просячы пытанне і Bayesians, *даследаванняў у галіне гісторыі і філасофіі навукі*, 30 (4), 687-697, 1999 год.

Байесовский, частотнай і класічных метадаў

Праблема з класічнага падыходу з'яўляецца тое, што з'яўляецца вынікам не аб'ектыўна абумоўленымі. Простае падзея адзін чалавек з'яўляецца складанае падзея іншага чалавека.Адзін даследчык можа спытаць, зноў адкрытая планета ", якая верагоднасць таго, што жыццё існуе на новай планеце?" у той час як іншы можа спытаць: "Якая верагоднасць, што вугляродныя існуе жыццё на ёй?"

Бруна Финетти, у прадмове да сваёй двухтомнай трактат аб байесовский ідэі, выразна гаворыцца, што "верагоднасці не існуе". Пад гэтым ён мае на ўвазе, што верагоднасць не знаходзяцца ў манеты або косці, яны не з'яўляюцца характарыстыкамі такіх рэчаў, як маса, шчыльнасць і г.д.

Некаторыя байесовского падыходу лічаць тэорыю верагоднасцяў як працяг дэдуктыўны логікі (у тым ліку дыялог логіку, пытальныя логікі, нефармальнай логікі і штучнага інтэлекту) для апрацоўкі нявызначанасці. Яно накіравана на тое, каб вывесці з першых прынцыпаў адназначна правільны спосаб прадстаўлення вашага прадстаўлення аб становішчы рэчаў, і іх абнаўлення ў святле доказаў. Законы верагоднасці маюць той жа статус, як і законы логікі. These байесовского падыходу відавочна "суб'ектыўны" у тым сэнсе, што яны маюць справу з праўдападобнасць якіх рацыянальны агент павінен прыкласці да прапановах ён/яна лічыць, "з улікам яго/яе цяперашні стан ведаў і досведу". У адрозненне ад гэтага, па меншай меры некаторыя не байесовского падыходу лічаць верагоднасці як "аб'ектыўныя" прыкметы рэчаў (або сітуацыі), якія на самай справе там (наяўнасць дадзеных).

Байесовский і класічная статыстыка аналіз тых жа дадзеных, як правіла дасягаюць такой жа выснова. Тым не менш, байесовский можа лепш даць колькасную ацэнку праўдзівай нявызначанасці ў сваім аналізе, асабліва пры істотных да інфармацыі. *Bayesians гатовыя прызначыць функцыі размеркавання верагоднасці (ы) з параметрам насельніцтва (ы), а не frequentists*.

З пункту гледжання навукоўца, ёсць усе падставы адмовіцца ад байесовского мыслення. Праблема ў тым, што байесовский здзелак развагі не аб'ектыўныя, а суб'ектыўныя верагоднасці. У выніку, любыя развагі выкарыстаннем байесовского падыходу не можа быць публічна праверыў - тое, што робіць яго, па сутнасці, бескарыснай для навукі, як і без эксперыментаў рэпликативной.

Байесовский перспектывы часта праліваюць карысны святло на класічных працэдур. Трэба ісці ў байесовской даць аснову для даверных інтэрвалаў імавернасны інтэрпрэтацыі якіх практыкуючыя часта хочуць, каб размясціць на іх. Гэта адкрыццё можа дапамагчы ў прыцягненні ўвагі да таго, што іншы апрыёрнае размеркаванне можа прывесці да розных інтэрвалам.

Байесовский могуць падмануць, засноўваючы папярэдняга размеркавання на дадзеных, частотнага можа засноўваць гіпотэзы павінны быць праверaны на гэтыя дадзеныя.Напрыклад, роля пратаколу клінічных выпрабаванняў, каб гэтага не адбывалася, патрабуючы ад гіпотэзы павінны быць указаны да збору дадзеных. Такім жа чынам, байесовский можа быць абавязаны пазначыць да грамадскасці ў пратакол да пачатку даследавання. У калектыўных навуковых даследаванняў, гэта будзе некалькі складаней, чым для частотнай гіпотэзы, таму што настаяцелі павінны быць асабістым для ўзгаднення правядзення.

Адпаведнае колькасць, якія былі прапанаваныя для ацэнкі нявызначанасці высноў, гэта значыць для апрацоўкі апрыёры нечаканым, верагоднасць таго, сама функцыя.

Калі вы выконваеце серыю аднолькавых выпадковых эксперыментаў (напрыклад, манета кідкоў), які ляжыць у аснове размеркавання верагоднасці таго, што максімізуе верагоднасць таго, што вынік вы назіралі гэта імавернасным размеркаванне прапарцыйна выніках эксперыменту.

Гэта мае прамое тлумачэнне таго, каб казаць, як (адносна) кожны з магчымых тлумачэнняў (мадэль), няхай гэта будзе атрыманае з дадзеных ці не, прадказвае назіраных дадзеных.Калі дадзеныя апынуцца крайнія ("атыповыя") у некаторым родзе, так што верагоднасць таго, паказвае на бедны набор мадэляў, гэта будзе хутка забраць у наступных раўндах навуковых даследаванняў з боку навуковай супольнасці. Не доўга гарантыі частата выканання, ні асабістага меркавання не патрабуецца.

Існуе, у якім сэнсе байесовский падыход арыентаваны на прыняцце рашэнняў і выпрабаванняў частотнай гіпотэзы падыход арыентаваны на навуку. Напрыклад, можа быць не дастаткова доказаў таго, што навукова агент X з'яўляецца шкодным для чалавека, але не можа быць апраўдана ў прыняцці рашэнняў, каб пазбегнуць гэтага ў рацыёне харчавання.

Амаль ва ўсіх выпадках, кропкавая адзнака з'яўляецца бесперапыннай выпадковай велічыні. Такім чынам, верагоднасць таго, што верагоднасць якой-небудзь канкрэтнай каштарысу Справа ў тым, сапраўды нуля. Гэта азначае, што ў вакууме інфармацыі, можна зрабіць здагадку аб тым, ці не верагоднасці. Нават калі ў нас ёсць інфармацыя, мы сапраўды можам толькі здагадвацца, дыяпазон верагоднасці.

Такім чынам, пры ацэнцы параметраў дадзенай групы насельніцтва, неабходна, каб кропкавая адзнака суправаджаецца некаторай ступені магчымыя памылкі ў ацэнцы. Шырока прымальным падыходам з'яўляецца тое, што кропкавая ацэнка павінна суправаджацца некаторым інтэрвалам аб ацэнцы з некаторай мерай забеспячэння, што гэты інтэрвал змяшчае праўдзівае значэнне параметру насельніцтва. Напрыклад, забеспячэння надзейнасці працэсаў у апрацоўчай прамысловасці заснаваны на дадзеных прыводзіцца інфармацыя для вырабу прадукту праектных рашэнняў.

Мэта байесовский: Існуе выразная сувязь паміж верагоднасцю і логіка: як уяўляецца, кажуць нам, як мы павінны разважаць. Але як, уласна, з'яўляюцца гэтыя два паняцці звязаны? Мэта Bayesians прапаноўвае адзін з адказаў на гэтае пытанне. У адпаведнасці з мэтай Bayesians, верагоднасць абагульняе дэдуктыўны логікі: дэдуктыўны логіка падказвае нам, якія высновы пэўныя, улічваючы мноства памяшканняў, у той час як верагоднасць кажа нам, у якой ступені варта верыць высновы, улічваючы памяшкання пэўныя высновы прысуджэннем поўнай меры веры. У адпаведнасці з мэтай Bayesians, памяшканняў аб'ектыўна (гэта значыць адназначна) вызначыць, у якой ступені варта верыць высновы.

Дадатковая літаратура : Бернарда Я., А. Сміт, байесовский тэорыі, М., 2000. Конгдон П., байесовского статыстычнага мадэлявання, М., 2001. Corfield Д. і Дж. Уільямсана, *асновы Bayesianism*, Kluwer Academic Publishers, 2001. Утрымоўвае логікі, матэматыкі, тэорыі прыняцця рашэнняў і крытыкі Bayesianism. зямлі Ф., *аператыўная Суб'ектыўныя статыстычныя метады*, М., 1996. Уяўляе сістэматычнае лячэнне субъективистской метады нараўне з добрым абмеркаваньнем гістарычных і філасофскіх фон з асноўных падыходаў да тэорыі верагоднасці і статыстыкі. Прэс С. *Суб'ектыўныя і аб'ектыўныя байесовский Статыстыка: прынцыпы, мадэлі і прыкладанні*. М.: Свет, 2002Цімермана Н., *тэорыі недакладных мностваў*, Kluwer Academic Publishers, 1991. Невыразнай логіцы падыходаў да верагоднасці (па Л. Задэ і яго паслядоўнікаў) уяўляюць розніцу паміж "магчымасць тэорыі" і тэорыі верагоднасцяў.

Слых, перакананняў, меркаванняў і фактаў

Статыстыка ёсць навука аб прыняцці рашэнняў ва ўмовах нявызначанасці, якія павінны быць заснаваныя на фактах, а не на чутках, асабістае меркаванне, ні па веры.

У неабходнасці рацыянальнага чалавечага стратэгічнага мыслення развіваліся, каб справіцца з яго/яе навакольнага асяроддзя. Рацыянальнае стратэгічнае мысленне, якое мы называем развагі яшчэ адзін сродак, каб зрабіць свет вылічаецца, прадказальны і больш кіраваным для утылітарных мэтаў. Пры пабудове мадэлі рэальнасці, фактычнай інфармацыі, такім чынам, неабходна для пачатку любога рацыянальнага стратэгічнага мыслення ў форме развагі. Тым не менш, мы не павінны блытаць факты перакананні, меркаванні або чуткі. У наступнай табліцы дапамагае растлумачыць адрозненні:

Слых, перакананняў, меркаванняў і фактаў

	Слых	**Вера**	**Меркаванне**	**Факт**
Адзін кажа сабе	Мне трэба выкарыстоўваць гэта ў любым выпадку	Гэта праўда. Я правоў	Гэта маё меркаванне	Гэта факт
Адзін кажа іншым	Гэта можа быць праўдай. Вы ведаеце!	Вы памыляецеся	Гэта значыць ваша	Я магу растлумачыць вам

Вераванні вызначаецца як уласнае разуменне нехта. У перакананні, "Я" заўсёды мае рацыю, і "вы" не правы.. Існуе нічога, што можна зрабіць, каб пераканаць чалавека, што яны лічаць няправільным
Што тычыцца веры, Анры Пуанкаре сказаў: "усё сумневы і верыць усяму, што: гэтыя дзве аднолькава зручныя стратэгіі, альбо мы абысціся без неабходнасці. думаю ". лічачы сродкаў, не жадаючы ведаць, што гэта факт. Чалавечыя істоты з'яўляюцца найбольш схільныя верыць таму, што яны менш за ўсё разумее. Такім чынам, вы можаце, а ёсць розум адкрыты дзіўна, чым адзін закрыты верай. Найбольшая засмучэнне розуму ў нешта верыць, таму што адзін хоча, каб гэта было так.

Гісторыя чалавецтва поўная трывожных перспектывах нарматыўна адлюстравана, напрыклад, інквізіцыя, паляванне на ведзьмаў, даносы, і прамывання мазгоў. "Святыя перакананні" не толькі ў рэлігіі, але і ў ідэалогіі, і можа нават ўключаць навукі. Шмат у чым такім жа чынам шматлікія навукоўцы спрабавалі "выратаваць тэорыю". Напрыклад, фрайдысцкі лячэнне з'яўляецца свайго роду прамывання мазгоў з боку псіхатэрапеўта, калі пацыент знаходзіцца ў навадны настрою цалкам і рэлігійнай веры ў любы тэрапеўт робіць яго/яе і вінаваціць сам/сама ва ўсіх выпадках. Існуе гэты велізарны нязграбны момант з часоў халоднай вайны, дзе мысленне яшчэ не ацанілі. Нішто так цвёрда верыў, як тое, што менш за ўсё вядома.

Гісторыя чалавецтва таксама ўсеяныя адмовіцца вера-мадэляў. Тым не менш, гэта не азначае, што той, хто не разумее, што адбываецца вынайшаў мадэль, ні не за камунальныя паслугі або практычнае значэнне. Асноўная ідэя складалася ў культурныя каштоўнасці любой іншай мадэлі. Памылковасць перакананні не абавязкова пярэчанні да веры. Пытанне ў тым, у якой ступені гэта жыццё садзейнічання і ўмацавання жыцця для верніка?

Меркаванні (або пачуцці) крыху менш экстрэмальных, чым перакананні, аднак яны з'яўляюцца дагматычных. Меркаванне азначае, што чалавек мае пэўныя ўяўленні, якія яны лічаць права. Акрамя таго, яны ведаюць, што іншыя маюць права на сваё ўласнае меркаванне. Людзі дачыненні да іншых меркаванняў і ў сваю чаргу, чакаюць таго ж. Пры фарміраванні свайго меркавання, эмпірычныя назіранні, відавочна, моцна залежыць ад стаўлення і ўспрымання. Тым не менш, думкі, якія добра ўкараніліся павінна расці і змяняцца, як здаровае дрэва. Факт толькі навучальны матэрыял, які можа быць прадстаўлены ў зусім не-дагматычных спосабам. Кожны мае права на яго/яе ўласным меркаванні, але ніхто не мае права памыляцца ў яго/яе фактамі.

Грамадская думка часта з'яўляецца свайго роду рэлігіяй, з большасцю, як яго прарока. Акрамя таго, прыбытак мае кароткую памяць і не забяспечвае паслядоўную думкі з цягам часу.

Чуткі і плёткі яшчэ слабей, чым меркаванне. Цяпер пытанне ў тым, хто будзе верыць гэтым? Напрыклад, чуткі і плёткі пра чалавека, гэта тыя, калі вы чуеце, што вам падабаецца, пра каго вы не робіце. Вось напрыклад, вы можаце быць знаёмыя з: Чаму няма Нобелеўскай прэміі па матэматыцы? Гэта *меркаванне* многіх, што Альфрэд Нобель злавіў сваю жонку ў любоўнай сітуацыі з Миттаг-Леффлера, перш за ўсё шведскі матэматык таго часу. Такім чынам, Нобелеўскі баяўся, што калі б ён стварыць матэматыку прэміі, першае, каб ён быў бы ML. Гісторыя паўтараецца, незалежна ад таго, як часта паўтарае просты *факт*, што Нобелеўскі не быў жанаты.

Каб зразумець розніцу паміж пачуццём і стратэгічнае мысленне, уважліва разгледзець наступныя дакладнае сцвярджэнне: Той, хто лічыць сябе самым шчаслівым чалавекам на самай справе так, але той, хто лічыць сябе мудрым, як правіла, найбольшы дурань. Большасць людзей не прасіце фактаў у прыняцці сваіх рашэнняў. Яны аддалі перавагу б адзін добры, душа-эмацыйнае задавальненне, чым дзясятка фактаў. Гэта не азначае, што вы не павінны адчуваць нічога. Звярніце ўвагу на свае пачуцці. Але не думаю, што з імі.

Факты адрозніваюцца ад перакананняў, чутак і меркаванняў. Факты з'яўляюцца асновай рашэнняў. Справа ў тым, што нешта правільна, а можна даказаць, каб быць праўдай на аснове фактычных дадзеных і лагічных аргументаў. Факт можа быць выкарыстаны, каб пераканаць сябе, сваіх сяброў, і ворагаў. Факты заўсёды могуць быць зменены. Дадзеныя становяцца інфармацыяй, калі яна становіцца стаўленне да вашага рашэння праблемы. Інфармацыя становіцца фактам, калі дадзеныя могуць падтрымаць яго. Факт становіцца веданнем, калі яно выкарыстоўваецца ў паспяховым завяршэнні структураваны працэс прыняцця рашэнняў. Тым не менш, факт становіцца меркаванне, калі яно дазваляе для розных інтэрпрэтацый, гэта значыць розныя пункты гледжання. Звярніце ўвагу, што здарылася ў мінулым, самай справе, не так. Ісціна гэта тое, што мы думаем пра тое, што адбылося (напрыклад, мадэлі).

Бізнес Статыстыка пабудавана на фактах, а дом камянямі. Але збор фактаў, не больш карысна і інструментальных навук для мэнэджара, чым куча камянёў дом.

Навука і рэлігія карэнным чынам адрозніваюцца. Рэлігія заклікае нас верыць без сумневу, нават (або асабліва) у сувязі з адсутнасцю надзейных доказаў. На самай справе, гэта вельмі важна для маюць веру. Навука патрабуе ад нас нічога не прымаць на веру, каб быць асцярожнымі нашай схільнасці да самападману, адмовіцца неафіцыйныя дадзеныя.Навука лічыць, глыбокі, але здаровы скептыцызм прэм'ер-функцыю. Адна з прычын яе поспеху ў тым, што навука мае ўбудаваны, выпраўляць памылкі машыны ў самым яе сэрцы.

Даведайцеся, як падысці да інфармацыі крытычна і адрозніваць прынцыповым чынам ад перакананні, меркаванні і факты. Крытычнае мысленне, неабходнае для вытворчасці аргументаваныя прадстаўлення рэальнасці ў працэсе мадэлявання. Аналітычнае мысленне патрабуе яснасці, паслядоўнасці, сведчыць, перш за ўсё, паслядоўнай, мэтанакіраванай, мыслення.

Дадатковая літаратура:
Boudon P., *Паходжанне значэння: сацыялогіі і філасофіі веры*, транзакцыі Publishers, London, 2001.
Кастанеда К., *Актыўная бок бясконцасці*, Harperperennial бібліятэка, 2000.
Гудвін П., Г. Райт, *рашэнні аналіз кіравання суда*, М., 1998.
Юр'евіч Р., *Падман фрэйдызм: вывучэнне амерыканскай прамывання мазгоў прафесіяналаў і аматараў*., Філадэльфія, Dorrance 1974
Каўфман В., *Рэлігія ў чатырох вымярэннях: экзыстэнцыйная і эстэтычная, гістарычная і параўнальная*, Ридерз Дайджэст Прэс, 1976.

Што такое Статыстычны аналіз дадзеных? Дадзеныя не інфармацыя!

Дадзеныя не з'яўляюцца інфармацыяй! Каб вызначыць, які статыстычны аналіз дадзеных, неабходна спачатку вызначыць статыстыку. Статыстыка ўяўляе сабой набор метадаў, якія выкарыстоўваюцца для збору, аналізу, прадставіць і інтэрпрэтаваць дадзеныя. Статыстычныя метады выкарыстоўваюцца ў самых розных прафесій і дапамагчы людзям выявіць, вывучыць і вырашыць шматлікія складаныя праблемы. У свеце эканомікі і бізнесу, гэтыя метады дазваляюць кіраўнікам і менеджэрам прымаць абгрунтаваныя і аптымальныя рашэнні адносна нявызначаных сітуацый.

Вялікая колькасць статыстычнай інфармацыі, даступных у глабальнай і сучасных эканамічных умовах з-за пастаяннага паляпшэння ў галіне камп'ютэрных тэхналогій. Каб паспяхова канкураваць у глабальным маштабе, кіраўнікі і асобы, якія прымаюць рашэнні павінны быць у стане зразумець інфармацыю і выкарыстоўваць яе эфектыўна.Статыстычны аналіз дадзеных забяспечвае практычны вопыт, каб заахвочваць выкарыстанне статыстычнага мыслення і метады прымяняюцца для таго, каб прымаць кампетэнтныя рашэнні ў дзелавым свеце.

Кампутары гуляюць вельмі важную ролю ў статыстычным аналізе дадзеных. Статыстычны пакет праграмнага забеспячэння, SPSS, якая выкарыстоўваецца ў гэтым, вядома, прапануе шырокі спектр апрацоўкі дадзеных і магчымасці шматлікіх статыстычных працэдур аналізу, які можа аналізаваць маленькіх да вельмі вялікіх статыстычных дадзеных.Кампутар будзе садзейнічаць абагульненню дадзеных, але статыстычны аналіз

дадзеных прысвечана інтэрпрэтацыі выходных, каб зрабіць высновы і прагнозы.

Вывучэнне праблемы з дапамогай статыстычнага аналізу дадзеных звычайна ўключае ў сябе чатыры асноўных этапы.

1. Вызначэнне праблемы
2. Збор дадзеных
3. Аналізуючы дадзеныя
4. Справаздачнасць аб выніках

Вызначэнне праблемы

Дакладнае вызначэнне праблемы неабходна для таго, каб атрымаць дакладныя дадзеныя пра яго. Вельмі цяжка збіраць дадзеныя без выразнага вызначэння праблемы.

Збор дадзеных

Мы жывем і працуем ва ўмовах, калі збор дадзеных і статыстычныя разлікі сталі лёгка амаль на мяжы пошласці. Парадаксальна, але дызайн збор дадзеных, не дастаткова падкрэсліваецца ў статыстычным аналізе дадзеных падручнікаў, былі паслабленыя бытующим меркаваннем, што шырокія вылічэнні могуць кампенсаваць любыя недахопы ў канструкцыі збору дадзеных. Трэба пачаць з акцэнтам на важнасць вызначэння насельніцтва, аб якой мы імкнемся, каб зрабіць высновы, усе патрабаванні адбору і эксперыментальнага праектавання павінны быць выкананы.

Распрацоўка спосабаў збору дадзеных з'яўляецца важнай задачай у галіне статыстычнага аналізу дадзеных. Два важных аспектаў статыстычнага даследавання:
насельніцтва - гэта сукупнасць усіх элементаў, цікавасць да даследавання
узораў - падмноства насельніцтва
Статыстычныя высновы ў ставіцца да пашырэння сваіх ведаў атрымаць ад выпадковай выбаркі з генеральнай сукупнасці для ўсяго насельніцтва. Гэта вядома ў матэматыцы як індуктыўных разваг. Гэта значыць, веды цэлага ад канкрэтнага. Яго асноўныя прыкладанні ў гіпотэз аб дадзенай групы насельніцтва.
мэта статыстычнага аналізу складаецца ў атрыманні інфармацыі аб інфармацыі аб насельніцтве, якія змяшчаюцца ў форме ўзору. Гэта проста не ўяўляецца магчымым праверыць ўсё насельніцтва, так што ўзор з'яўляецца адзіным рэальным спосабам атрымання дадзеных з-за часу і кошту. Дадзеныя могуць быць колькаснымі або якаснымі. Якасныя дадзеныя пазнак ці імёнаў, якія выкарыстоўваюцца для вызначэння атрыбутаў кожнага элемента. Колькасныя дадзеныя заўсёды лічбавы і паведамляць, колькі і як шмат.

Для статыстычнага аналізу дадзеных, адрозненні паміж перасекам і часовых шэрагаў дадзеных мае вялікае значэнне. Крыжаваныя дадзеныя сабраныя дадзеныя паўторна на тым жа або прыкладна такой жа момант часу. Дадзеных часовых шэрагаў дадзеных, сабраных за некалькі перыядаў часу.

Дадзеныя могуць быць сабраны з існуючых крыніц або атрыманы з дапамогай назіранняў і эксперыментальных даследаванняў, накіраваных на атрыманне новых дадзеных. У эксперыментальным даследаванні, зменная цікавасць выяўлена. Тады адзін або больш фактараў у даследаванні кантралююцца, так што дадзеныя могуць быць атрыманы аб тым, як фактараў, якія ўплываюць на зменныя. У наглядальных даследаваннях, не робіцца ніякіх спробаў кантраляваць або ўплываць на зменныя, якія прадстаўляюць інтарэс.Апытанне з'яўляецца, бадай, найбольш распаўсюджаны тып обсервационное даследаванне.

Аналіз дадзеных

Статыстычны аналіз дадзеных дзеліць метады для аналізу дадзеных на дзве катэгорыі: пошукавыя метады і пацвярджаюць метадаў. Пошукавыя метады выкарыстоўваюцца, каб выявіць, што дадзеныя, здаецца, гаворыць з дапамогай простай арыфметыкі і простых ў маляваць карцінкі для абагульнення дадзеных. Якія пацвярджаюць метады выкарыстоўваюць ідэі тэорыі верагоднасцяў ў спробе адказаць на канкрэтныя пытанні. Верагоднасць гуляе важную ролю ў прыняцці рашэнняў, паколькі ён забяспечвае механізм для вымярэння, выразы, а таксама аналіз нявызначанасцяў, звязаных з будучымі падзеямі. Большасць тым, закранутых у дадзеным курсе падпадаюць пад гэтую рубрыку.

Справаздачнасць аб выніках

Праз высновы, ацэнка або тэст заявы аб характарыстыках насельніцтва можа быць атрыманы з ўзору. Вынікі могуць быць прадстаўлены ў выглядзе табліц, графікаў або набор працэнты. Таму што толькі невялікая калекцыя (узор) не былі агледжаны і не за ўсё насельніцтва, атрыманыя вынікі павінны адлюстроўваць нявызначанасць з дапамогай верагоднасці заявы і інтэрвалы значэнняў.

У заключэнне, найважнейшым аспектам кіравання любой арганізацыі плануюць у будучыні. Разумны сэнс, інтуіцыю і разуменне стану эканомікі можа даць мэнэджару прыкладныя прадстаўленне або "пачуццё", што можа адбыцца ў будучыні. Тым не менш, ператвараючы гэта пачуццё ў нумар, які можна эфектыўна выкарыстоўваць цяжка.Статыстычны аналіз дадзеных дапамагае менеджэрам прагназаваць і прадказваць будучыню аспекты вядзення бізнесу. Найбольш паспяховыя менеджэры і кіраўнікі тыя, хто можа зразумець інфармацыю і выкарыстоўваць яе эфектыўна.

Наведайце таксама [розныя падыходы да статыстычнага мыслення]

Апрацоўка дадзеных: кадаваньне, увод і рэдагаванне

Дадзеныя часта рэгіструюцца ўручную на дадзеныя ліста. Калі лік назіранняў і зменных невялікі дадзеныя павінны быць прааналізаваны на кампутары. Гэтыя дадзеныя будуць затым прайсці праз тры этапы:

Кадаванне: дадзеныя перадаюцца, у выпадку неабходнасці закадаваныя лістоў.

Увод: дадзеныя ўводзяцца і захоўваюцца па крайняй меры, два незалежных ўводу дадзеных асоб. Напрыклад, калі Бягучы абследаванне насельніцтва і іншых штомесячных абследаванняў былі прынятыя з выкарыстаннем папяровых апытальнік, Бюро перапісу насельніцтва ЗША выкарыстоўвалі спараныя клавіша ўводу дадзеных.

Рэдагаванне: дадзеныя правяраюцца шляхам параўнання дзвюх незалежных тыпаў дадзеных. Стандартная практыка для ключавых ўводу дадзеных з папяровых анкет, каб ўвесці ўсе дадзеныя ў два разы. У ідэале, у другі раз павінна быць зроблена іншым ключавым аператарам, чыя праца ўступленне ў прыватнасці, уключае праверку неадпаведнасці паміж арыгінальнай і 2. Запісы. Лічыцца, што гэта "double-key/verification" метад дае 99,8% дакладнасць хуткасці для поўнага націску клавіш.

Тып памылкі: Запіс памылкі, памылкі друку, памылкі транскрыпцыі (няправільнае капіяванне), інверсія (напрыклад, 123,45 набіраецца як 123,54), паўтарэнне (калі лік паўтараецца), наўмыснае памылку.

Тып дадзеных і ўзроўні вымярэння

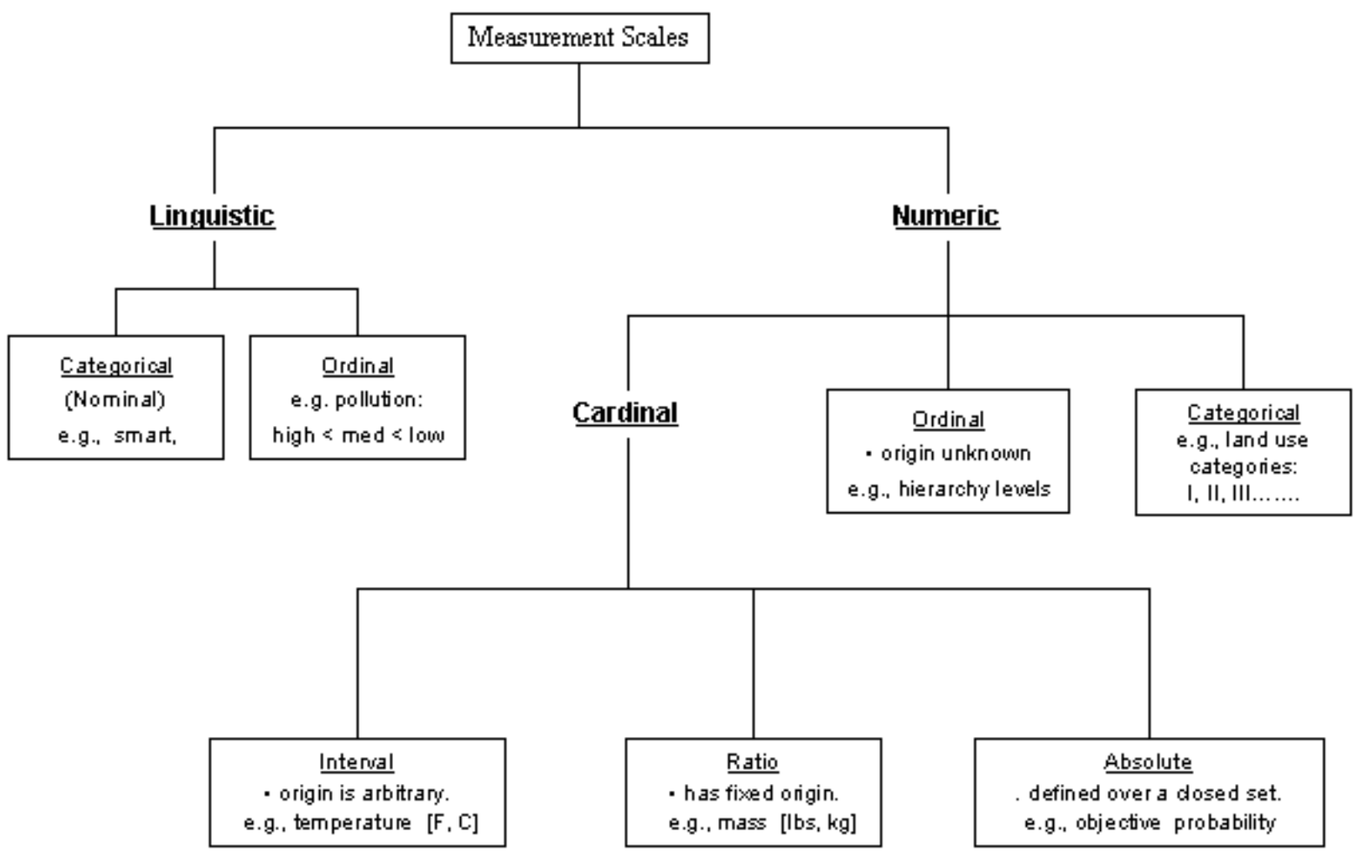
Інфармацыя можа быць сабрана ў галіне статыстыкі з выкарыстаннем якасных і колькасных дадзеных.

Якасныя дадзеныя, такія як колер вачэй групай асоб, якія не вычислимых арыфметычных адносін. Гэта пазнакі, якія раяць, да якой катэгорыі або класе чалавека, аб'ект або працэс падзення. Яны называюцца катэгарыяльныя зменныя.

Колькасныя набораў дадзеных складаецца з мер, якія прымаюць лікавыя значэнні, для якіх апісання, такія як сярэднія і стандартныя адхіленні маюць сэнс. Яны могуць быць уведзеныя ў парадку і падзеленыя на дзве групы: дыскрэтныя дадзеныя або бесперапынных дадзеных. Дыскрэтных дадзеных падліковых дадзеных, напрыклад, колькасць дэфектных вырабаў, вырабленых у працэсе вытворчасці кожны дзень. Бесперапынныя дадзеныя, калі параметры (зменныя) вымерна, выяўляецца ў бесперапыннай шкале.Напрыклад, вымярэння вышыні асобы.

Першым мерапрыемствам у галіне статыстыкі складаецца ў вымярэнні ці лічыць. Вымярэнне/кошт тэорыя мае справу з сувязі паміж дадзенымі і рэальнасцю. набор дадзеных з'яўляецца прадстаўленне (напрыклад, мадэль) у рэальнасці на аснове лікавых і рытмічны маштабах. Дадзеныя званы "першасны тып" дадзеных, калі аналітык прымаў удзел у зборы дадзеных, якія маюць дачыненне да яго/яе расследавання. У адваротным выпадку, ён называецца "другасным" дадзеных.

Дадзеныя прыходзяць у форме намінальнай, парадкавай, інтэрвал і суадносіны (успомніце французскія NOIR слова чорны колер). Дадзеныя могуць быць бесперапынным або дыскрэтным.



Measurement Scales

Абодва нуля і адзінкі вымярэння ў адвольнай інтэрвальнай шкале. У той час як адзінкай вымярэння з'яўляецца адвольным ў шкале адносін, яе нулявая кропка з'яўляецца натуральным атрыбутам. Катэгарыяльнай зменнай вымяраецца ў парадкавай або намінальнай шкале.

Вымярэнне тэорыя мае справу з сувязі паміж дадзенымі і рэальнасцю. Абодва статыстычнай тэорыі і тэорыі вымярэнняў неабходна, каб зрабіць высновы аб рэальнасці.

З статыстыкі жыць дакладнасці, яны аддаюць перавагу інтэрвал/Суадносіны узроўняў вымярэння.

Праблемы з Паэтапнае пераменная выбару

Вось некаторыя з распаўсюджаных праблем са ступеністым зменнай адбору ў Рэгрэсійная аналізу.

1. Гэта дае R-квадрат значэння, якія моцна прадузятае высокая.
2. F і хі-квадрат тэсты цытуе побач з кожнай зменнай на раздрукоўцы не сцвярджаў размеркавання.
3. Метад дае даверныя інтэрвалы для эфектаў і прагнозных значэнняў, якія ілжыва вузкім.
4. Гэта дае P-значэння, якія не маюць дакладнага значэння і правільнае выпраўленне для іх з'яўляецца вельмі складанай праблемай
5. Гэта дае прадузятае каэфіцыентаў рэгрэсіі, якія павінны ўсаджванне, т. е. каэфіцыенты для астатніх зменных з'яўляюцца занадта вялікімі.
6. Ён мае сур'ёзныя праблемы пры наяўнасці коллинеарности.
7. Ён заснаваны на метадах (напрыклад, F-тэст для укладзеных мадэляў), якія павінны былі быць выкарыстаны для праверкі загадзя абумоўленага гіпотэз.
8. Павелічэнне памеру пробы не вельмі дапамагала.

Адзначым таксама, што ўсё магчыма-падмноства падыход не прыводзіць да выдалення з пералічаных вышэй праблем.

Дадатковая літаратура:
Дерксен, С. і Г. Кесельман, назад, наперад і паэтапна аўтаматызаваныя алгарытмы выбару падмноства, *British Journal матэматычных і статыстычных псіхалогіі*, 45, 265-282, 1992.

Альтэрнатыўны падыход для ацэнкі лініі рэгрэсіі

Наступны падыход з'яўляецца так званы "непараметрический метад" для ацэнкі параметраў у просты рэгрэсіі y = TX+ Ь:

1. Перапішыце у = ТХ+ Ь а б =-т+ у.
2. Кожная кропка дадзеных (х я, у мяне) адпавядае лініі б =-х я м у+ я ў декартовой плоскасці каардынат (т, бы), і ацэнка м і б могуць быць атрыманы ад перасячэння пар такіх ліній. Ёсць больш п (п+1)/2 такіх ацэнак.
3. Вазьміце медыяны, каб атрымаць канчатковыя адзнакі.

Дадатковая літаратура:
Корниш-Боуден А. *Аналіз кінэтычных дадзеных ферментаў*, Oxford Univ Press, 1995.
Hald A., *Гісторыя матэматычнай статыстыкі: з 1750 па 1930 год*, М., Нью-Ёрк, 1998 год. Сярод іншага, аўтар адзначае, што ў пачатку 18-га стагоддзя даследаванні былі чатыры розныя метады для вырашэння праблемы ўстаноўкі: Майер-Лапласа метадам сярэдніх, Бошкович-Лапласа метадам найменшых абсалютных адхіленняў, метад Лапласа мінімізацыі найбуйнейшых абсалютная рэшткавы і Лежандра метад мінімізацыі сумы квадратаў рэшткаў. Толькі адзін спосаб выбару паміж гэтымі метадамі было: параўнаць вынікі ацэнак і рэшткаў.

Шматмернага аналізу дадзеных

Дадзеныя лёгка збіраць, што мы сапраўды маем патрэбу ў комплекснае рашэнне праблемы інфармацыі. Мы можам прагледзець базу дадзеных як вобласць, якая патрабуе датчыкі і прылады для вымання неабходнай інфармацыі. Як і ў вымярэнні працэсу, адпаведныя інструменты развагі павінны быць ужытыя да задачы інтэрпрэтацыі дадзеных. Эфектыўныя інструменты выступаюць у двух якасцях: для абагульнення дадзеных і аказанне дапамогі ў інтэрпрэтацыі. Мэта тлумачэння сродкі павінны раскрываць дадзеныя на некалькіх узроўнях дэталізацыі.

Даследаванне недакладных дадзеных выявы часам патрабуецца шырокавугольны аб'ектыў, каб паглядзець у яго паўнаце. У іншых выпадках патрабуецца буйным планам аб'ектыў, каб засяродзіцца на дробных дэталях. Графічна інструменты, якія мы выкарыстоўваем забяспечыць такую гнуткасць. Большасць хімічных сістэм з'яўляюцца складанымі, паколькі яны звязаны многімі зменнымі і ёсць шмат узаемадзеянняў паміж зменнымі. Такім чынам, хемометрические метады абапіраюцца на шматмерных статыстычных і матэматычных інструментаў, каб раскрыць ўзаемадзеянне і паменшыць памернасць дадзеных.

Шматфактарнага аналізу з'яўляецца філіялам статыстыкі, уключаючы разгляд аб'ектаў на кожным з якіх назіраюцца значэння ліку зменных. Шматмерныя метады выкарыстоўваюцца ва ўсім дыяпазоне палёў статыстычнага прымянення: у медыцыне, фізічных і біялагічных навук, эканомікі і сацыяльных навук, і, вядома, у многіх прамысловых і камерцыйных прыкладанняў.

Аналіз асноўных кампанент выкарыстоўваецца для вывучэння дадзеных для памяншэння памеру. Як правіла, СПС спрабуе прадстаўляць п. карэлявалі выпадковых велічынь паменшаны набор некоррелированных зменныя, якія атрымліваюцца шляхам пераўтварэння зыходнага набору на адпаведныя подпространства. Некоррелированных зменныя выбіраюцца добрымі лінейнай камбінацыяй зыходных зменных, з пункту гледжання тлумачэння максімальнай дысперсіяй, артаганальных напрамках у дадзеных. Два цесна звязаных з імі метадаў, аналіз галоўных кампанент і фактарнага аналіз, выкарыстоўваюцца для зніжэння памернасці шматмерных дадзеных. У гэтых метадаў карэляцыі і ўзаемадзеяння паміж зменнымі прыведзены ва ўмовах невялікай колькасці асноўных фактараў. Метады хуткага вызначэння ключавых пераменных або груп зменных, якія кіруюць вывучаемай сістэмы. У выніку паніжэння памернасці дазваляе таксама графічнае паданне дадзеных, так што значныя сувязі паміж назіраннямі або узораў можа быць ідэнтыфікаваны.

Іншыя метады ўключаюць у сябе мнагамернае шкалирование, кластарны аналіз і Аналіз адпаведнасцяў.

Дадатковая літаратура:
Чатфилд С., А. Колінз, *Уводзіны ў шматмерны аналіз*., Чэпмен і Хол, 1980
. Хойл Р., Статыстычная стратэгіі для малых даследаванняў узораў, Thousand Oaks, Каліфорнія, Sage, 1999
Krzanowski В., *Прынцыпы шматмерных Аналіз пункту гледжання карыстача*, М.: Свет, 1988.
Mardia К. Дж. Кента і Дж. Бибби, *шматмернага аналізу*, М., 1979.

Значэнне і тлумачэнне Р-значэння (тое, што дадзеныя кажуць?)

Р-значэнне, якое напрамую залежыць ад канкрэтнага асобніка, спрабуе забяспечыць пэўную сілу вынікаў тэстаў, у адрозненне ад проста адхіліць або не адхіліць. Калі нулявая гіпотэза дакладная і верагоднасць выпадковых змяненняў з'яўляецца адзінай прычынай прыклад адрозненні, то Р-значэнне з'яўляецца колькаснай мерай карміць ў працэсе прыняцця рашэнняў у якасці доказаў. У наступнай табліцы прыведзены разумнай інтэрпрэтацыі Р-значэння:

Р-значэнне	Інтэрпрэтацыя
Р < 0,01	вельмі моцныя сведчанні супраць Н0
0,01 £ Р < 0,05	умераным доказы супраць Н0
0,05 £ Р < 0,10	навадны доказаў супраць Н0
0,10 £ Р	мала ці ўвогуле няма ніякіх рэальных доказаў супраць Н0

Гэтая інтэрпрэтацыя атрымала шырокае прызнанне, і многія навуковыя часопісы рэгулярна публікуюць артыкулы з дапамогай гэтай інтэрпрэтацыі выніку праверкі гіпотэз.

Для фіксаванага памеру выбаркі, калі колькасць рэалізацый вызначаецца загадзя, размеркаванне р раўнамерна (пры ўмове нулявой гіпотэзы). Мы хацелі б выказаць гэта ў якасці Р (р £ х) = х. Гэта азначае, што крытэрый р < 0,05 дасягае 0,05.

Пры р-значэнне звязана з наборам дадзеных, з'яўляецца мерай верагоднасці таго, што дадзеныя маглі паўстаць як выпадковую выбарку з некаторай папуляцыі апісваецца статыстычнай (тэставанне) мадэлі.

Р-значэнне з'яўляецца мерай таго, наколькі ў вас ёсць доказы супраць нулявой гіпотэзы. Чым менш р-значэнне, тым больш доказаў ў вас ёсць. Можна спалучаць р-значэнне з узроўнем значнасці для прыняцця рашэнняў па дадзенай праверкі гіпотэз. У такім выпадку, калі значэнне р менш некаторага парогавага значэння (звычайна 0,05, часам трохі больш як 0,1 ці крыху менш як 0,01), то адмовіцца ад нулявой гіпотэзы.

Зразумейце, што размеркаванне р-значэння пры нулявой гіпотэзы Н0 раўнамерна, і, такім чынам, не залежыць ад канкрэтнага віду статыстычных тэстаў. У статыстычны тэст гіпотэзы, значэнне Р з'яўляецца верагоднасць выяўлення тэставай статыстыкі па крайняй меры, як крайняя як значэнне сапраўды назіраецца, пры ўмове, што нулявая гіпотэза дакладная. Значэнне р вызначаецца па адносінах да размеркавання. Такім чынам, мы маглі б назваць яго "мадэлі размеркавання гіпотэза", чым "нулявая гіпотэза".

Карацей кажучы, гэта проста азначае, што калі нулявая была праўда, значэнні р верагоднасць супраць нулявой ў гэтым выпадку. Р-значэнне вызначаецца назіранае значэнне, аднак, гэта абцяжарвае нават сцвярджаць зваротнае р.

Вы можаце як з дапамогай Р-значэння для папулярных дыстрыбутываў аплетаў Java.

Дадатковая літаратура:
Аршам Х., Койпера Р-значэнне, сродкаў вымярэння і працэдура для дабра, згоды, выпрабаванняў, *чаconic Applied Statistics*, Vol. 15, № 3, 131-135, 1988.

Дакладнасць, дакладнасць, надзейнасць і якасць

Дакладнасць ставіцца да блізкасці вымярэння "фактычнага" і "рэчаіснасць" значэнне фізічнай велічыні, у той час як тэрмін дакладнасці выкарыстоўваецца, каб паказаць блізкасць, з якой вымярэнняў адпавядаюць адзін з адным зусім незалежна ад сістэматычнай хібнасці ўдзел. Такім чынам, "дакладныя" адзнака мае невялікі ўхіл. "Дакладны" адзнака мае малыя зрушэння і дысперсіі. Якасць прапарцыйна зваротнай дысперсіі.

Надзейнасць працэдуры, у якой ступені яго ўласцівасці не залежаць ад тых здагадак, якія вы не хочаце рабіць. Гэта мадыфікацыя арыгінальнай версіі Box, і гэта ўключае ў сябе байесовский меркаванняў, страты, а таксама раней. Цэнтральная лімітавая тэарэма (ЦПТ) і Гаўса-Маркава кваліфікаваць як ўстойлівасць тэарэмы, але Huber-Hempel вызначэнне не падпадае пад катэгорыю надзейнасці тэарэмы.

Мы заўсёды павінны адрозніваць надзейнасці і эфектыўнасці зрушэння надзейнасці. Здаецца, для мяне відавочна, што ні статыстычная працэдура можа быць надзейным ва ўсіх сэнсах. Трэба быць больш канкрэтнымі аб тым, што працэдура павінна быць абаронена ад. Калі выбарачнае сярэдняе часам разглядаецца як надзейная ацэнка, гэта таму, што ЦПТ гарантуе 0 зрушэння пры вялікіх выбарках, незалежна ад зыходнага размеркавання. Гэтая адзнака з'яўляецца зрушэнне надзейныя, але гэта відавочна не надзейны, як і эфектыўнасць яе дысперсія можа павялічыцца да бясконцасці. Гэта адхіленне можа быць нават бясконцым, калі асноўны размеркавання Коши або Парэта з вялікім параметрам маштабу. Гэта прычына, па якой выбарачнае сярэдняе не хапае надзейнасці ў адпаведнасці з Huber-Хампель азначэнні. Праблема ў тым, што М-ацэнкі выступаюць Хубер, Хампель і некалькі іншых людзей, з'яўляецца зрушэнне надзейнай, толькі калі зыходнае размеркаванне сіметрычна.

У рамках выбарачнага абследавання, два тыпу статыстычных высноў маюцца: мадэль на аснове вывадаў і распрацоўка на аснове вывадаў, які выкарыстоўвае толькі рандомизации, спалучаныя з працэсам выбаркі (без здагадкі аб неабходнасці мадэлі). Аб'ектыўнае дызайн на аснове ацэнкі, як правіла, называюць надзейнай ацэнкі, таму што несмещенности дакладна для ўсіх магчымых размеркаванняў. Здаецца відавочным, аднак, што гэтыя ацэнкі ўсё яшчэ можа быць дрэннай якасці, як дысперсія, якая можа быць празмерна вялікім.

Тым не менш, іншыя людзі будуць выкарыстоўваць гэтае слова ў іншым (недакладны) спосабамі. Том Кендал. 2 Сучасная тэорыя статыстыкі, а таксама прыводзіць Box 1953, і ён робіць менш карысным заявы аб здагадках. Акрамя таго, Кендал дзяржаў у адным месцы, надзейнасці сродкаў (толькі) аб тым, што тэст памеру, застаецца пастаяннай пры розных умовах. Гэта тое, што людзі выкарыстоўваюць, па-відаць, калі яны сцвярджаюць, што двухбаковае Т-тэстаў "надзейны", нават калі адхіленні і памеры выбаркі не роўныя.Асабіста я не люблю называць тэсты надзейнай, калі дзве версіі Т-тэст, які прыкладна ў роўнай ступені надзейнымі, могуць мець 90% розныя вынікі пры параўнанні узораў якіх трапляюць у інтэрвал адмовы (або рэгіёну).

Мне лягчэй выкарыстоўваць фразу: "Існуе розніца надзейны", што азначае, што той жа выснова прыходзіць незалежна ад таго, як вы выканайце тэст, што (апраўданай) пераўтварэнне вы выкарыстоўваеце, калі вы падзяляеце ацэнкі для праверкі на дыхатаміі і г.д., або тое, што знешніх уплываў вы трымаеце пастаянную якасці ковариант.

Функцыя ўплыву і яе прыкладання

Уплыў функцыі ацэнкі ў кропцы х, па сутнасці, змены ў ацэнках, калі бясконца малых назіранняў дадаюцца ў кропцы х, дзеленай на масу назіранняў. Уплыў функцыя дае бясконца малой адчувальнасці рашэння да дадання новай сістэме каардынат.

Гэта асноўны патэнцыял прымянення функцыі ўплыву ў параўнанні метадаў ацэнкі для ранжыравання надзейнасці. Разумны сэнс формы ўплыву функцыю надзейнай працэдуры, калі экстрэмальныя значэння адкідаюцца, гэта значыць дадзеныя аздабленнем.

Ёсць некалькі асноўных статыстычных тэстаў, такіх як тэст на выпадковасць, крытэрыі аднастайнасці насельніцтва, тэст для выяўлення планавальнік (ы), а затым тэст на нармалёвасць. Для ўсіх гэтых неабходных выпрабаванняў маюцца магутныя працэдуры статыстычнага аналізу дадзеных літаратуры. Акрамя таго, паколькі аўтары абмяжоўваюць свае ўяўленні тэстаў сярэдняй, яны могуць выклікаць ЦПТ для, скажам, любы ўзор памерам больш за 30.

Канцэпцыя ўплыву з'яўляецца вывучэнне ўплыву на высновы і высновы ў розных абласцях даследаванняў, у тым ліку статыстычнага аналізу дадзеных. Гэта магчыма за кошт абурэнняў аналізу. Напрыклад, функцыя ўплыву ацэнкі з'яўляецца змяненне ў ацэнцы, калі бясконца малая змяненне ў адным назіранні, дзялення на колькасць змяненняў. Ён дзейнічае як аналіз адчувальнасці ацэнкі.

Уплыў функцыі былі пашыраныя, каб "што-калі" аналізу, надзейнасці і аналізу сцэнарыяў, такіх як даданне ці выдаленне назірання, outliners (ы) ўздзеяння, і гэтак далей.Напрыклад, для дадзенага размеркавання як нармальных, так ці інакш, для якіх насельніцтва параметры былі ацэненыя з узорамі, даверны інтэрвал для ацэнкі сярэдняга або сярэдняга менш, чым для тых значэнняў, якія маюць тэндэнцыю да канечнасці, такія як 90% або 10 % дадзеных. У той час як пры ацэнцы сярэднім на можа выклікаць цэнтральнай лімітавай тэарэмы для любой выбарцы аб'ёму па параўнанні, скажам 30. Тым не менш, мы не можам быць упэўнены, што разліковы розніца ёсць праўдзівая розніца насельніцтва і, такім чынам, большай нявызначанасцю, поўзае і адзін трэба падаць у суд на функцыі ўплыву ў якасці вымяральнага інструмента працэдуры прыняцця рашэнняў.

Дадатковая літаратура:
Мельнікаў І., *Уплыў функцыі і матрыцы*, Dekker, 1999.

Што такое недакладнай верагоднасці?

Недакладныя верагоднасць з'яўляецца агульным тэрмінам для многіх матэматычных мадэляў, якія вымяраюць выпадкова ці нявызначанасці, без рэзкіх колькаснага верагоднасцяў.Гэтыя мадэлі ўключаюць у сябе веру функцый, тэорыі магчымасцяў, параўнальнай спарадкаванасці верагоднасці, выпуклыя мноства імавернасны мер, недакладных мер, інтэрвал значэннямі верагоднасці, магчымасці мер, дакладнасці мер, а таксама верхняя і ніжняя чакання або прадбачання. Такія мадэлі неабходныя для высновы праблемы, дзе адпаведнай інфармацыі не хапае, нявызначаныя або супярэчлівыя, а таксама ў вырашэнні праблем, дзе перавагу можа быць няпоўным.

Што такое мета-аналіз?

Мета-аналіз разглядае набор вынікаў, каб даць агульны вынік, які носіць ўсёабдымны характар і сілу.

а) Асабліва, калі эфект, памеры досыць малыя, ёсць надзея, што можна атрымаць добрую ўладу, па сутнасці робячы выгляд, што маюць больш N, як у сіле, камбінаваная выбарка.

б) калі велічыня эфекту дастаткова вялікае, то дадатковая магутнасць не патрабуецца для асноўных эфектаў дызайну: Замест гэтага, ён тэарэтычна можа быць можна паглядзець на кантраст паміж невялікімі зменамі ў працах саміх сябе.

Напрыклад, каб параўнаць дзве велічыні эфекту (г), атрыманыя з двух асобных даследаванняў, вы можаце выкарыстаць:

Z = (Z 1 - Z 2)/[(1/п 1 -3)+ (1/п 2 -3)] 1/2

дзе Z 1 і Z 2 з'яўляюцца Фішэр пераўтварэнні г, і два н я аўтара ў назоўніку ўяўляюць памер выбаркі для кожнага даследаванні.

Калі вы сапраўды ўпэўненыя, што "пры іншых роўных умовах" пройдзе ўверх. Тыповы "мета"-даследаванне не рабіць тэсты на аднастайнасць, што павінны быць абавязаны

Іншымі словамі:

1. ёсць аб'ём даследаванняў/літаратурныя дадзеныя, якія вы хацелі абагульніць

2. Складаецца разам усе дапушчальныя прыклады літаратуры (нататка: некаторыя з іх могуць быць адкінутыя па розных прычынах)

3. некаторыя дэталі кожнага расследавання расшыфроўваюцца... Найважнейшае значэнне мела б пра тое, што быў ці не быў знойдзены, г.зн., наколькі больш у адзінках SD з'яўляецца выкананне групе лячэння ў параўнанні з адным ці некалькімі элементамі кіравання.

4. называюць значэння ў кожным з даследаванняў у # 3.. Памеры міні-эфект.

5. усіх дапушчальных наборах дадзеных, вы спрабуеце падвесці агульны памер эфекту за кошт фарміравання набору асобных эфектаў... і выкарыстанне агульнай памяці SD ў якасці дзельніка.. Такім чынам саступаючы па сутнасці сярэдні памер эфекту.

6. У літаратуры аналіз мета... Часам гэтыя велічыні эфекту дадаткова пазначаныя як малых, сярэдніх ці вялікіх....

Вы можаце паглядзець на эфект памеры па-рознаму.. розных фактараў і зменных., Але ў двух словах, гэта тое, што зроблена.

Я ўспамінаю выпадак у фізіцы, у якой пасля з'ява назіралася ў паветры, эмульсіі дадзеныя былі разгледжаны. Тэорыя будзе мець каля 9% эфекту ў эмульсіі, і вось, апублікаваныя дадзеныя далі 15%. Як гэта часта бывае, не было ніякіх істотных адрозненняў (практычныя, а не статыстычнай) у тэорыі, а таксама не памылка ў дадзеных. Гэта было толькі тое, што вынікі эксперыментаў, у якіх нічога статыстычна значным аказалася не паведамлялася.

Гэта непаведамленне падобных эксперыментаў, а часцяком і канкрэтныя вынікі, якія не былі статыстычна значнымі, які ўводзіць асноўныя прадузятасці. Гэта таксама спалучаецца з цалкам памылковае стаўленне даследчыкаў, статыстычна значныя вынікі з'яўляюцца важнымі, і, чым калі б гэта не мае значэння, эфект быў не важны. Нам сапраўды трэба адрозніваць тэрмін "статыстычна значнае", а звычайныя словы значным.

Мета-аналіз з'яўляецца спрэчным тыпам агляд літаратуры, у якіх вынікі асобных рандомізіраванный кантраляваных даследаванняў, якія аб'яднаны разам, каб паспрабаваць атрымаць ацэнку ўплыву ўмяшання вывучаюцца. Гэта павялічвае статыстычную магутнасць і выкарыстоўваюцца для вырашэння праблемы справаздачы, якія не згодныя адзін з адным. Гэта не проста добра, і ёсць шмат прыроджаных праблем.

Дадатковая літаратура:
Липси М., Д. Уілсан, *практычныя мета-аналіз*, Sage Publications, 2000.

Якое ўплыў памеру

Памер эфекту (ES) уяўляе сабой стаўленне сярэдняй розніцай у стандартнае адхіленне, гэта значыць форма Z-рахунак. Выкажам здагадку, што эксперыментальная група лячэння мае сярэдні бал Xe і кантрольная група мае сярэдні бал Xc і стандартнае адхіленне навук, то велічыня эфекту роўная (Xe - Xc)/Sc

Эфект памеру дазваляе параўнальнае ўплыў розных метадаў лячэння ў параўнанні нават тады, калі на аснове розных узораў і розныя вымяральныя прыборы.

Такім чынам, ES-сярэдняя розніца паміж кантрольнай групай і групай лячэння. Howevere, метадам Гласс, ES з'яўляецца (mean1 - mean2)/SD з кантрольнай групы, у той час як на метад Хантэр-Шміт, у ES ёсць (mean1 - mean2)/SD аб'яднання і рэгулюецца інструмент каэфіцыент надзейнасці. ES звычайна выкарыстоўваецца ў мета-аналіз і аналіз магутнасці.

Дадатковая літаратура:
Купер Х. і Л. Hedges,*Даведнік Сінтэз даследаванняў*, Нью-Ёрк, Russell Sage, 1994.
Липси М., Д. Уілсан, *практычныя мета-аналіз*, Sage Publications, 2000.

Што такое закон Бенфорда? А як наконт закона Ципфа?

Што такое закон Бенфорда: закон Бенфорда абвяшчае, што калі мы выпадковым чынам выбраць нумар з табліцы фізічных канстант або статыстычныя дадзеныя, верагоднасць таго, што першая лічба будзе "1", складае каля 0,301, а не 0,1, як мы маглі б чакаць, калі ўсе лічбы аднолькава верагодныя. Увогуле, "закон" кажа, што верагоднасць таго, што першая лічба "D":

$$P\{d\} = \frac{\ln\left(1+\frac{1}{d}\right)}{\ln(10)}$$

Гэта азначае, што лік у табліцы фізічных канстант, хутчэй за ўсё, пачынаюцца з лічбы менш, чым больш лічба. Гэта можна назіраць, напрыклад, шляхам вывучэння табліцы лагарыфмаў і адзначаючы, што на першых старонках значна больш зношаных і недакладнае, чым пазней старонак.

Метады скарачэння зрушэння

Найбольш эфектыўнымі прыладамі для зрушэння скарачэнне не з'яўляецца прадузятай ацэнкі з'яўляюцца пачатковай загрузкі і Jackknifing.

Паводле легенды, барон Мюнхгаўзен выратаваўся ад патапленьня ў зыбучыя пясках, пацягнуўшы за сябе, выкарыстоўваючы толькі свае валасы. Статыстычныя пачатковай загрузкі, які выкарыстоўвае передискретизации з зададзенага набору дадзеных, каб імітаваць зменлівасць, якая прывяла да дадзеных у першую чаргу, мае значна больш надзейны тэарэтычныя асновы і можа быць вельмі эфектыўнай працэдурай для ацэнкі хібнасці велічыні ў статыстычных задач.

Bootstrap з'яўляецца стварэнне віртуальнага насельніцтва шляхам дублявання і таго ж ўзору зноў і зноў, а затым паўторна ўзоры з віртуальнага насельніцтва, каб сфармаваць набор спасылак. Затым параўнайце вашыя арыгінальныя ўзоры са спасылкай ўсталяваць, каб атрымаць дакладнае значэнне р. Вельмі часта, пэўныя структуры "на сябе", так што рэшткавы разлічваецца для кожнага выпадку. Тое, што затым зноў пробы ад мноства рэшткаў, якія затым дадаюцца ў тых структурах мяркуецца, да некаторай статыстыкай ацэньваецца. Мэтай часта для ацэнкі Р-ўзроўню.

Складанага нажа гэта паўторна вылічыць дадзеныя, пакідаючы на назіранне кожны раз. Пакінуць адзін-з рэплікацыі дае вам тыя ж тэматычныя ацэнкі, я думаю, у якасці належнага складаны нож ацэнкі. Jackknifing робіць некалькі лагічных складаныя (адкуль "складанага нажа" - паглядзець яго), каб забяспечыць ацэнкі каэфіцыентаў і памылак, што (вы спадзеяцеся) будзе паменшаны ўхіл.

Метады зрушэння скарачэнне мець шырокае прымяненне ў антрапалогіі, хіміі, кліматалогіі, клінічныя выпрабаванні, кібернетыкі і экалогіі.

Дадатковая літаратура:
Эфрон Б., *складанага нажа, загрузкі і іншыя Передискретизация планы*., Сіям, Філадэльфія, 1982
Эфрон Б. і Р. Tibshirani, *увядзенне ў пачатковай загрузкі*, Chapman & Hall (у цяперашні час CRC Press), 1994.
Шао Ј., Д. аў, *складанага нажа і пачатковай загрузкі*, Springer Verlag, 1995.

Плошчу пад стандартнай нармальнай крывой

Прыблізны плошчу пад стандартнай нармальнай крывой ад 0 да Z з'яўляецца

```
Z (4.4-Z)/10 для 0 £ Z £ 2,2   0,49 2,2 < Z < 2,6     0,50 Z £ 2,6   максімальная абсалютная памылка вышэй набліжэнні  прыкладна палова
```

Колькасць класаў у інтэрвал гістаграмы

Перш чым мы зможам пабудаваць наша размеркаванне частот мы павінны вызначыць, колькі класаў мы павінны выкарыстоўваць. Гэта чыста адвольна, але занадта мала класаў або занадта шмат класаў не будзе забяспечваць максімальна дакладную карціну, як можа быць атрымана з яшчэ амаль аптымальнае колькасць. Эмпірычная залежнасць (вядомая як правіла Sturges "), які, як правесці і якія могуць быць выкарыстаны ў якасці кіраўніцтва па ліку класаў (к) даецца

да = найменшае цэлае, большая ці роўнае 1+ Log (N)/Log (2) = 1+ 3.332Log (п)

Для таго, каб "аптымальны" Вы маеце патрэбу ў некаторай мерай якасці - верагодна, у дадзеным выпадку, "лепшы" спосаб адлюстравання любы існуючы аб'ём інфармацыі ў дадзеных. Памер выбаркі спрыяе гэтаму, так што звычайныя прынцыпы павінны выкарыстоўваць ад 5 да 15 класаў, адным трэба больш класаў, калі вам даводзіцца вельмі вялікая выбарка. Вы прымаеце пад увагу перавагі акуратна шырыні класа, пажадана кратна 5 або 10, таму што гэта лягчэй ацаніць маштаб.

Акрамя гэтага яна становіцца прадметам меркаванні - апрабаваць шэраг класе шырыню і выбраць той, які працуе лепш за ўсё. (Мяркуецца, у вас ёсць кампутар і можа генераваць альтэрнатыўныя гістаграмы даволі лёгка).

Ёсць часта пытанні кіравання, якія ўваходзяць у яго. Напрыклад, калі вашы дадзеныя ў параўнанні з аналагічнымі дадзенымі - такія, як папярэднія даследаванні, або з іншых краін - вы абмежаваныя інтэрвалам у ім выкарыстоўваліся.

Калі гістаграма вельмі скажоныя, то няроўныя класы павінны быць разгледжаны. Выкарыстанне вузкіх класаў, дзе клас высокіх частот, шырокія класы, дзе яны нізкія.

Наступныя падыходы з'яўляюцца агульнымі:

Хай п выбаркі, то лік класаў інтэрвал можа быць

MIN {п ½, 10log (п)}.

Такім чынам, для 200 назіранняў вы будзеце выкарыстоўваць 14 інтэрвалаў, але ў 2000 годзе вы будзеце выкарыстоўваць 33.

Акрамя таго,

1. Знайсці дыяпазону (максімальнае значэнне - мінімальнае значэнне).

2. Падзяліце дыяпазон на разумны памер інтэрвалу: 2, 3, 5, 10 або некалькім = 10.

3. Імкнецеся да не менш чым 5 інтэрвалаў і не больш за 15.

Мадэляванне структурнымі раўнаннямі

Структурныя метады мадэлявання ўраўненні выкарыстоўваюцца для вывучэння сувязяў паміж зменнымі. Адносіны, як правіла, лічыцца лінейным. У сацыяльных і паводніцкіх даследаванняў, большасць з'яў пад уплывам вялікага ліку дэтэрмінант, якія звычайна маюць складаную структуру узаемасувязяў. Для таго каб зразумець адносную важнасць гэтых дэтэрмінант іх адносіны павінны быць адэкватна прадстаўлены ў мадэлі, якая можа быць зроблена з мадэлявання структурнымі раўнаннямі.

Структурная мадэль ўраўненні могуць прымяняцца да адной групе выпадкаў або некалькіх груп выпадкаў. Калі некалькі груп, аналізуюцца параметры могуць быць абмежаваныя роўнымі паміж двума або больш групамі. Калі дзве або больш групы, аналізуюцца, значыць, назіраецца і на схаваныя зменныя таксама могуць быць уключаны ў мадэль.

У якасці дадатку, як вы праверыць роўнасць рэгрэсіі схілы з таго ж прыкладу з выкарыстаннем 3-х розных метадаў вымярэння? Вы можаце выкарыстоўваць структурны падыход да мадэлявання.

1 - стандартызацыя ўсіх трох набораў дадзеных да аналізу, паколькі б вагі таксама залежыць ад дысперсіі предиктор пераменных і стандартызацыі, вы выдаліце гэты крыніца.

2 - мадэль залежнай пераменнай, як эфект ад усіх трох мер і атрымаць шлях каэфіцыент (б вага) для кожнага з іх.

3 - Затым ўсталюйце мадэль, у якой тры каэфіцыента шляху абмежаваныя роўнымі. Калі значнае памяншэнне ў патрэбнае месца, шляхі не роўныя.

Дадатковая літаратура:
Schumacker P., P. Ломакс, *кіраўніцтва для пачаткоўцаў, каб Мадэляванне структурнымі раўнаннямі*, Лоўрэнс Erlbaum, Нью-Джэрсі, 1996 год.

Эканометрики і мадэляў часавых шэрагаў

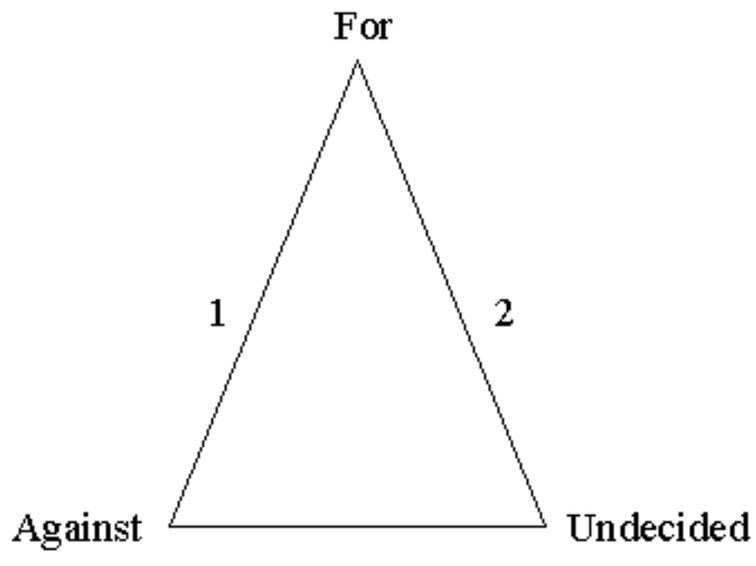
Эканометрики мадэлі набораў адначасова рэгрэсіі з прыкладаннямі ў такіх галінах, як эканоміка прамысловасці, сельскай гаспадаркі і карпаратыўнай стратэгіі і рэгулявання. Мадэлі часовых шэрагаў патрабуе вялікай колькасці назіранняў (скажам больш 50). Абедзве мадэлі паспяхова выкарыстоўваюцца для бізнес-прыкладанняў, пачынаючы ад мікра да макра даследаванні, уключаючы фінансы і эндагеннага росту. Іншыя падыходы да мадэлявання ўключаюць структурныя і класічныя мадэлі, такія як Харві, і бокс-Джэнкінс падыходы, з-аналізу і інтэграцыі агульных мікра эканометрики ў імавернасны мадэлях, напрыклад, логит, прабіты і Товита, панэльных дадзеных і перасекаў. Эканометрика, у асноўным, вывучэннем пытання аб прычыннасці, гэта значыць пытанне аб выяўленні прычынна-следчай сувязі паміж вынікам і набор фактараў, якія могуць быць вызначаны такога зыходу.У прыватнасці, ці рэалізацыі гэтай канцэпцыі ў ў часовых шэрагах, і экзагенных мадэлявання.

Дадатковая літаратура:
Ericsson M. і Дж. Айронс, тэставанне экзагеннага, Oxford University Press, 1994.
С. Грэнджэр, П. Newbold, прагназавання ў бізнесе і эканоміцы, М., 1989.
Хамудам А. Дж. Роўлі, (рэдакцыя), мадэляў часавых шэрагаў, прычыны і экзагенных, Эдвард Элгар Pub., 1999.

Tri-лінейныя каардынаты трыкутніка

"Трайны схеме" звычайна выкарыстоўваецца, каб паказаць змена думкі (FOR - СУПРАЦЬ - не вызначыліся). Трохкутнай дыяграме выкарыстоўваюцца ў першую чаргу па хімік Ўілард Гібс ў сваіх даследаваннях на фазавых пераходах. Ён заснаваны на здагадцы, што з геаметрыяй ў выглядзе роўнабаковага трыкутніка, сума адлегласцяў ад любой кропкі да трох бакоў пастаянна. Гэта азначае, што працэнт складу з сумесі трох рэчываў можа быць прадстаўлена як кропка ў такой схеме, так як сума працэнтаў з'яўляецца пастаяннай (100). Тры вяршыні з'яўляюцца кропкамі чыстых рэчываў.

Тое ж самае справядліва і для "кампазіцыі" з меркаванняў насельніцтва. Калі працэнты за, супраць і не вызначыліся суму да 100, тую ж тэхніку для прадстаўлення могуць быць выкарыстаныя. Глядзіце дыяграму ніжэй, якія варта разглядаць з непрапарцыйным ліст. Праўда роўнабаковага не могуць быць захаваны ў перадачы. Напрыклад, няхай першапачатковы склад меркаванняў быць прадастаўлена 1. Гэта значыць, некалькі не вызначыліся, прыкладна ў роўнай ступені як для супраць. Дазвольце іншым складам зададзена кропка 2. Гэтая кропка ўяўляе сабой больш высокі адсотак не вызначыўся і сярод вырашыў, большасць "за".



Унутраныя і Inter-ацэншчык надзейнасць

"Унутраная надзейнасць" шкалы часта вымяраецца каэфіцыент Кронбаха. Гэта актуальна, калі вы будзеце вылічыць агульны бал і вы хочаце ведаць, яго надзейнасць, заснаваная на няма іншага рэйтынгу. "Надзейнасць" ў * адзнака * ад сярэдняга карэляцыі, і ад колькасці элементаў, так як больш маштаб будзе (як мяркуецца) будзе больш надзейным. Лі дэталі маюць тыя ж сродкі як правіла, не важна.

Таў-эквівалентныя: сапраўдныя ацэнкі па пунктах, як мяркуецца, адрозніваюцца адзін ад аднаго не больш чым на сталай. У роўнай надзейнасці меры, якія складаюць яго элементы павінны быць па крайняй меры таў-эквівалентныя, калі гэта здагадка не выконваецца, гэта ніжняя мяжа ацэнкі надзейнасці.

Роднасныя меры: Гэта найменш абмежавальных мадэль у рамках класічнай тэорыі тэстаў патрабуе толькі, што сапраўднае балаў па мерах Кажуць, што вымярэння і таго ж з'явы выдатна карэлююць. Такім чынам, на роднасныя меры дысперсіі памылак, праўда, кошт сродкаў, а сапраўднае кошт адхіленні могуць быць няроўнымі

Для "Інтэр-ацэншчык" надзейнасць, адно адрозненне ў тым, што значэнне ляжыць на надзейнасць адзінага рэйтынгу. Выкажам здагадку, мы маем наступныя дадзеныя

```
Удзельнікі час Q1 Q2 Q3 на Q17  001 1 май 4 красавік 4  002 1 4 Сакавік 3 сакавіка  001 2 4 Красавік 3 мая  і г.д.
```

Вывучаючы дадзеныя, я думаю, ніхто не можа зрабіць лепш, чым глядзець на парны Т-тэст і Pearson карэляцыі паміж кожнай парай рэйтынгавых агенцтваў - т-тэст скажа вам ці сродкі розныя, у той час як карэляцыя кажа вам ці суды ў адваротным выпадку паслядоўна.

У адрозненне ад Пірсан, "ўнутры класа" суадносіны мяркуе, што рэйтынгавыя агенцтва сапраўды маюць той жа сярэдняе. Гэта не дрэнна, як і агульнае рэзюмэ, і гэта менавіта тое, што некаторыя рэдактары хочуць бачыць прадстаўлены на надзейнасць ўсёй рэйтынгавых агенцтваў. Гэта і плюс і мінус, што існуе некалькі розных формул для унутры-класа суадносіны, у залежнасці ад надзейнасці якіх ацэньваецца.

Для такіх мэтаў, як планаванне харчавання для прапанаванага даследавання, яно незалежна ад таго, рэйтынгавыя агенцтвы, якія будуць выкарыстоўвацца будзе ў дакладнасці тыя ж асобы. Добрая метадалогія прымянення ў такіх выпадках, з'яўляецца Bland & Altman аналізу.

SPSS каманды:

```
Надзейнасць (Alpha, KR-20) НАДЗЕЙНАСЦЬ
```

SAS каманды:

```
Надзейнасць (Alpha, KR-20) кар ALPHA
```

Калі варта выкарыстоўваць непараметрические тэхнікі?

Параметрычныя метады з'яўляюцца больш карыснымі, чым больш вы ведаеце аб сваім прадмеце, так як веданне вашага прадмета могуць быць ўбудаваны ў параметрычнай мадэлі. Непараметрические метады, у тым ліку абодвух сэнсах гэтага слова, распаўсюджванне бясплатнае тэставанне і гнуткія функцыянальныя формы, больш карысныя, тым менш вы ведаеце аб сваім прадмеце. Трэба выкарыстоўваць статыстычны метад, званы непараметрический, калі яна задавальняе па крайняй меры, з наступных пяці тыпаў крытэрыяў:

1. Уводу дадзеных аналізу пералічэння - гэта значыць, лічыць дадзеныя, якія прадстаўляюць колькасць назіранняў у кожнай катэгорыі або крос-катэгорыі.

2. Дадзеныя вымяраюцца і/або прааналізаваны з выкарыстаннем намінальнай шкалы вымярэння.

3. Дадзеныя вымяраюцца і/або прааналізаваны з дапамогай парадкавай шкалы вымярэння.

4. Выснова не тычыцца параметру ў размеркаванні насельніцтва - як, напрыклад, гіпотэза аб тым, што час-спарадкаванае мноства назіранняў паказвае выпадковым чынам.

5. Размеркавання верагоднасцяў статыстыкі, на якім аналіз заснаваны не залежыць ад канкрэтнай інфармацыі або здагадкі аб насельніцтве (ы), ўзору (ов) ўзятыя, але толькі на агульных здагадках, такіх як бесперапыннае і/або сіметрычным Размеркаванне насельніцтва.

Згодна з гэтым азначэнні, адрозненне непараметрических прадастаўлены альбо з-за ўзроўню вымярэння, якія выкарыстоўваюцца або неабходныя для аналізу, а ў тыпах з 1 па 3, тып высновы, а ў тыпе 4 ці агульнасці здагадак аб насельніцтве размеркавання, а ў 5 тыпу.

Напрыклад, можна выкарыстоўваць Манна-Уітні ранг выпрабаванняў як непараметрические альтэрнатывы студэнтаў Т-тэст, калі адзін не мае нармальнае размеркаванне дадзеных.

Манна-Уітні: Для выкарыстання з двума незалежнымі групамі (па аналогіі з незалежнымі групамі т-тэст)
Вілкоксона: Для выкарыстання з двух звязаных (напрыклад, адпаведнасць або паўторнае) груп (па аналогіі з адпаведнымі ўзорамі Т-тэст)
Kruskall-Уоліс: Для выкарыстання з двума або больш незалежнымі групамі (па аналогіі з адным фактарам паміж суб'ектамі-ANOVA)
Фрыдман: Для выкарыстання з двума або больш роднасных груп (па аналогіі з адным фактарам унутры суб'ектаў ANOVA)

Аналіз няпоўных дадзеных

Метады, якія займаюцца аналізам дадзеных з прапушчанымі значэннямі можна падзяліць на:

- Аналіз выпадкаў поўнай, у тым ліку ўзважванне карэктывы,
- Чарговасць метады і пашырэння некалькіх абвінавачванне, і
- метады, якія аналізуюць няпоўныя дадзеныя непасрэдна, не патрабуючы прастакутны набор дадзеных, такіх як максімальнага праўдападобнасьці і байесовских метадаў.

Некалькі абвінавацілі (IM) з'яўляецца агульнай парадыгмы для аналізу няпоўных дадзеных. Кожны адсутнічае дадзенае замяняецца м > 1 мадэлюецца каштоўнасці, вырабляючы м мадэлявання версіі поўныя дадзеныя. Кожная версія аналізуецца стандартнай камплектацыі, дадзеныя метады і вынікі аб'ядноўваюцца з дапамогай простых правіл, каб вырабіць высноў заявы, якія ўключаюць адсутнічае нявызначанасць дадзеных. Асноўны ўпор робіцца на практыцы MI для рэальных статыстычных задач сучаснай вылічальнай асяроддзя.

Дадатковая літаратура:
Рубіна Д. *Некалькі Чарговасць для ўлонне у апытаннях*, Нью-Ёрк, М., 1987.
Schafer J., *аналізу шматмерных дадзеных Няпоўныя*, Лондан, Чэпмен і Хол, 1997.
Маленькая Р., Д. Рубін, *статыстычны аналіз з адсутнічаюць дадзенымі*, Нью-Ёрк, М., 1987.

Ўзаемадзеяння ў ANOVA і Рэгрэсійная аналізу

Узаемадзеянне ігнаруюцца, толькі калі вы дазволіце гэта. Па гістарычных прычынах, ANOVA праграмы, як правіла вырабляць усе магчымыя ўзаемадзеяння, у той час (некалькі) рэгрэсіі праграмы звычайна не вырабляюць ніякіх узаемадзеянняў - па крайняй меры, не так рэгулярна. Так што да карыстачу пабудаваць ўзаемадзеянне ўмовах, калі з выкарыстаннем Рэгрэсійная аналізу праблемы, калі ўзаемадзеяння, ці, можа быць, цікавасць. (Пад "ўмовы ўзаемадзеяння" Я маю на ўвазе зменныя, якія ажыццяўляюць інфармацыйнае ўзаемадзеянне, уключаныя ў якасці прадказальнікаў ў Рэгрэсійная мадэлі.)

Рэгрэсіі з'яўляецца адзнака ўмоўнае матэматычнае чаканне выпадковай велічыні гэтага іншы (магчыма, вектар-) выпадковай велічыні.

Самы просты канструкцыі, каб перамнажаць прадказальнікаў, узаемадзеянне якіх павінна быць уключана. Пры наяўнасці больш чым у тры прадказальнікаў, і асабліва, калі сыравіну зменныя прымаюць значэння, далёкія ад нуля (напрыклад, колькасць элементаў справа), розных вырабаў (для шматлікіх узаемадзеянняў, якія могуць быць атрыманы), як правіла, цесна звязаны адзін з сябрам і з арыгінальнай предикторов. Гэта часам называюць "праблемай мультиколлинеарности", хоць было б дакладней назваць ілжывым мультиколлинеарности. Гэта магчыма, і часта рэкамендуецца, каб наладзіць сыравіны, каб зрабіць іх артаганальных да зыходных пераменным (і малодшыя члены ўзаемадзеяння, а).

Што значыць, калі стандартны тэрмін памылкі высокая? Multicolinearity гэта не адзіны фактар, які можа прывесці да вялікі ў SE для ацэнак "крутасці" каэфіцыентаў рэгрэсіі любых мадэляў. SE, назад прапарцыйныя дыяпазон зменлівасці ў предиктором зменнай. Напрыклад, калі вы былі ацэнкі лінейнай сувязі паміж вагой (x) і некаторыя дихотомических вынікаў і x = (50,50,50,50,51,51,53,55,60,62) SE будзе нашмат больш, чым калі x = (10,20,30,40,50,60,70,80,90,100) пры іншых роўных умовах. Існуе ўрок для планавання эксперыментаў. Для падвышэння дакладнасці ацэнкі, павялічыць дыяпазон ўваходных дадзеных. Яшчэ адной прычынай вялікага ў SE з'яўляецца невялікая колькасць "падзея" назірання або невялікая колькасць "не-падзеі" назірання (па аналогіі з малой дысперсіяй ў выніку зменная). Гэта не строга кантраляваны, але павялічыць ўсе ацэнкі SE (не толькі асобныя SE). Існуе і іншая прычына высокага ўзроўню памылак, гэта называецца серыйнай карэляцыі. Гэтая праблема часта, калі не тыповая, пры выкарыстанні часовых шэрагаў, бо ў гэтым выпадку выпадковыя тэрмін парушэнні часта адлюстроўваюць зменныя, не ўключаныя ў відавочным выглядзе ў мадэлі, якія могуць змяняцца павольна, як час праходзіць міма.

У лінейнай мадэлі, якая прадстаўляе змена залежнай пераменнай Y ў выглядзе лінейнай функцыі некалькіх незалежных зменных, узаемадзеянне паміж дзвюма незалежнымі зменнымі X і W можна прадставіць сваю прадукцыю, гэта значыць пераменная створана шляхам памнажэння іх разам. Алгебраічных такая мадэль прадстаўлена:

Y =+ B1x+ b2+ b3 W XW+ e.

Калі X і W з'яўляюцца катэгорыі сістэм. Гэта раўнанне апісвае 2 дысперсійнай аналізу (новая) мадэль, калі X і W, (квазі) бесперапынных зменных гэта раўнанне апісвае множественной лінейнай рэгрэсіі (МЛР) мадэлі.

Ва ўмовах новая, пра існаванне ўзаемадзеянне можа быць апісана як розніца паміж адрозненні: розніца ў сродках паміж дзвюма ўзроўнямі X на адно значэнне W не з'яўляецца такой жа, як розніца ў адпаведныя сродкі на іншае значэнне W, і гэтую не-жа-Несс ўяўляе сабой узаемадзеянне паміж X і W, яна колькасна значэннем b3.

Ва ўмовах MLR, узаемадзеянне прадугледжвае змяненне вугла нахілу (ад рэгрэсіі Y на X) ад аднаго значэння да іншага W значэнне W (ці, эквівалентна, змяненне нахілу рэгрэсіі Y на W для розных Значэнні X): у двух-прадказальнік рэгрэсіі з узаемадзеяннем паверхні водгуку гэта не плоскасць, а кручаная паверхню (напрыклад, "выгнуты волава печыва", у (1990 Дарлінгтона) фразу). Змяненне нахілу колькасна значэннем b3. Для вырашэння гэтай праблемы некалькі коллинеарности.

Дысперсія нелінейнай выпадковых функцый

Змена нелінейнай функцыі некалькіх выпадковых велічынь можа быць набліжана да "дэльта метад". Прыблізна дисперсия для гладкай функцыі F (X, Y) двух выпадковых велічынь (X, Y) атрымліваецца апраксімуецца F (X, Y) па лінейнай пункту гледжання Тэйлара ў наваколлі аб узоры сродкі X і Y.
Напрыклад, дысперсія XY і X/Y на аснове вялікага памеру выбаркі апраксімуецца:

[E (Y)] 2 Var (X)+ [E (X)] 2 Var (Y)+ 2 E (X) E (Y) Cov (X, Y)

і

Var (X)/([E (Y)] 2)+ Var (Y) ([E (X)] 2)/([E (Y)] 4) - 2 Cov (X, Y) E (X)/([E (Y)] 3)

адпаведна.

Візуалізацыя Статыстыка: Аналітычная геаметрыя, і статыстыка

Уводзіны ў візуалізацыі статыстыкі

Большая частка статыстычнай апрацоўкі даных уключае ў сябе Алгебраічныя аперацыі ў наборы дадзеных. Аднак, калі набор даных змяшчае больш чым на 3 нумары, што немагчыма ўявіць сабе яе геаметрычнае прадстаўленне, у асноўным за кошт чалавечых сэнсарных абмежаванняў. Геаметрыя мае значна больш даўнюю гісторыю, чым алгебру.Старажытныя грэкі прыкладной геаметрыі для *вымярэння зямлі*, а таксама распрацаваны *геаметрычныя* мадэлі. *Аналітычнай геаметрыі*, каб знайсці *эквівалентнасць между алгебрай і геаметрыяй*. Мэтай з'яўляецца лепшае разуменне візуалізацыі ў 2-х ці 3-мерным прасторы, а таксама абагульніць ідэі для больш высокіх памерам аналітычным мысленнем.

Без абмежавання супольнасці, і эканоміі месца, наступныя прэзентацыі ў рамках невялікага памеру выбаркі, што дазваляе нам бачыць статыстыку ў 1 або 2-мерным прасторы.

Сярэднюю і медыяна

Выкажам здагадку, што чатырох чалавек хочуць сабрацца разам, каб гуляць у покер. Яны жывуць на 1 -й вуліцы, 3 -я вуліца, 7 -й вуліцы і 15 -й стрыт. Яны хочуць, каб выбраць дом, які ўключае ў сябе мінімальная колькасць кіравання для ўсіх зацікаўленых бакоў.

Давайце выкажам здагадку, што яны вырашылі звесці да мінімуму абсалютнага колькасці руху. Калі б яны сустрэліся ў 1 -й вуліцы, колькасць кіраванне будзе 0+ 2+ 6+ 14 = 22 блокаў. Калі б яны сустрэліся на 3 -й вуліцы, колькасць ваджэння будзе 2+ 0+ 4+ 12 = 18 блокаў. Калі б яны сустрэліся ў 7 -й вуліцы, 6+ 4+ 0+ 8 = 18 блокаў. Нарэшце, на 15 -йстрыт, 14+ 12+ 8+ 0 = 34 блокаў.

Такім чынам, два дамы, што дасць магчымасць звесці да мінімуму колькасць кіраванне будзе 3 -й ці 7 -й вуліцы. На самай справе, калі яны хочуць нейтральны сайт, у любым месцы на 4 -й, 5 -й або 6 -й вуліцы таксама будзе працаваць.

Звярніце ўвагу, што любое значэнне ад 3 да 7 можа быць вызначана як сярэдняя 1, 3, 7 і 15. Такім чынам, сярэдняя з'яўляецца значэннем, якое зводзіць да мінімуму абсалютная адлегласць да кропак дадзеных.

Цяпер, чалавек, на 15 -й засмучаны на заўсёды прыходзіцца рабіць больш кіравання. Такім чынам, група згодная разгледзець розныя правілы. Пры прыняцці рашэння, каб мінімізаваць квадрат адлегласці кіравання, мы выкарыстоўваем прынцып найменшых квадратаў. Па ўзвядзення ў квадрат, мы даем больш вагі на адзін вельмі працяглай язды, чым на кучу кароткіх паездак. Пры гэтым правілы, 7 - й вуліцы дом (36+ 16+ 0+ 64 = 116 квадратных

блокаў) пераважней 3 -й вуліцы дом (4+ 0+ 16+ 144 = 164 квадратных блокаў). Калі вы лічыце, у любым месцы, а не толькі саміх дамоў, то 9 -й вуліцы месца, якое мінімізуе квадрат адлегласці прывадам.

Знайсці значэнне x, якое зводзіць да мінімуму:

$$(1 - x) 2+ (3 - x) 2+ (7 - x) 2+ (15 - x) 2.$$

Значэнне, якое зводзіць да мінімуму сумы квадратаў значэнняў 6.5, якая таксама роўная сярэдняе арыфметычнае 1, 3, 7 і 15. Пры вылічэнні, лёгка паказаць, што гэта мае месца ў цэлым.

Разгледзім невялікі прыклад рахункі з цотных лікам выпадкаў, напрыклад, 1, 2, 4, 7, 10 і 12. Медыяна 5,5, сярэдзіна інтэрвалу паміж дзесяткамі 4 і 7.

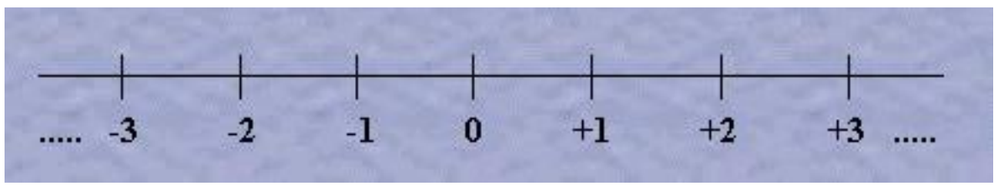
Як мы ўжо казалі вышэй, гэта праўда, што сярэдняя кропка, вакол якой сума абсалютных адхіленняў зведзены да мінімуму. У нашым прыкладзе сума абсалютных адхіленняў 22.Тым не менш, гэта не адзіная кропка. Любая кропка ў 4-х да 7 рэгіён будзе мець такое ж значэнне 22 на суму абсалютнага адхіленні.

Сапраўды, медыяны з'яўляюцца складаным. 50% вышэй - 50% ніжэй, не зусім правільна. Напрыклад, 1, 1, 1, 1, 1, 1, 8 не мае сярэдняе. Канвенцыя кажа, што медыяна 1, аднак каля 14% дадзеных ляжаць строга над ёй, 100% дадзеных, якія больш ці роўная сярэдняй.

Мы будзем выкарыстоўваць гэтую ідэю ў Рэгрэсійная аналізу. У аналагічны аргумент, лініі рэгрэсіі з'яўляецца унікальнай лініяй, якая мінімізуе суму квадратаў адхіленняў ад яго.Існуе не унікальная лінія, якая зводзіць да мінімуму сума абсалютных адхіленняў ад яго.

Арыфметычным і сярэднім геаметрычным

Арыфметычная: Выкажам здагадку, у вас дзве кропкі x і y, на рэальнае колькасць лініі восі:

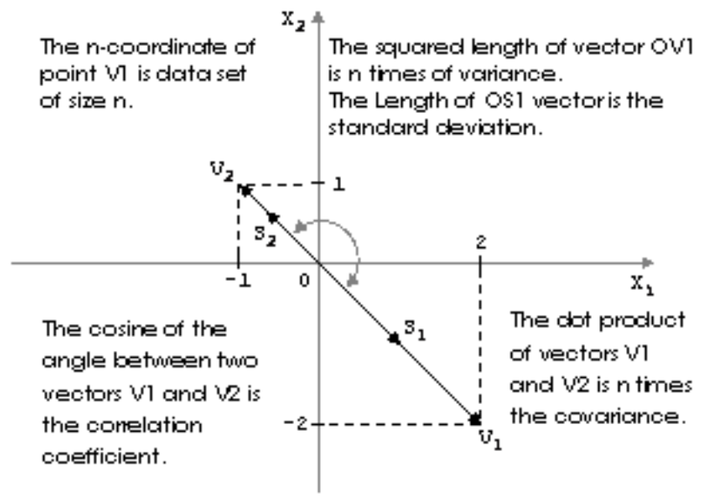


Сярэдняе арыфметычнае (a) такая кропка, што наступныя **вектарныя адносіны** сцвярджэнняў: бык - оа = габарытныя - ай.

Геаметрычнае сярэднюю: Выкажам здагадку, у вас ёсць два пазітыўных дадзеных кропак x і y, на вышэй рэальная колькасць лініі восі, то геаметрычнае сярэднюю (г) гэтых лікаў з'яўляецца кропка г такая, што | бык |/| ог | = | ог |/| ай |, дзе | бык | азначае **даўжыню адрэзка** бык, напрыклад.

Адхіленне, ковариация і каэфіцыент карэляцыі

Разгледзім набор дадзеных, які змяшчае п = 2 назірання (5, 1). Пасля цэнтралізацыі дадзеных, можна атрымаць вектар V1 = (5/3 = 2, 1-3 = -2), як паказана ў наступным п = 2 мернай сістэме каардынат:



Analytic-Geometry Representation of Major Statistics

Звярніце ўвагу, што даўжыня вектару V1 з'яўляецца:

$$| V1 | = [(2) 2+ (-2) 2] ½ = 8 ½$$

Дысперсія V1 з'яўляецца:

$$Var (V1) = S X я 2 /п = | V1 | 2 /п = 4$$

Стандартнае адхіленне:

$$| OS1 | = | V1 |/п ½ = 8 ½ /2 ½ = 2.$$

Зараз разгледзім 2. Назірання (2, 4). Акрамя таго, ён можа быць прадстаўлены вектар V2 = (-1, 1).

Ковариации,

$$Coy (V1, V2) = скалярны твор/п = [(2) (-1)+ (2) (1)]/2 = -4/2 = -2$$

Таму:

$$н Coy (V1, V2) = скалярны твор двух вектараў V1 і V2$$

Звярніце ўвагу, што інтэрнэт-прадукт множання двух даўжынь разы косінус вугла паміж двума вектарамі. Такім чынам,

$$Coy (V1, V2) = | OS1 | ' | OS2 | ' Cos (V1, V2) = (2) (1) Cos (180 °) = -2$$

Каэфіцыент карэляцыі складае:

$$p = Cos (V1, V2)$$

Магчыма, гэта найпростае доказ таго, што каэфіцыент карэляцыі заўсёды абмежаваны інтэрвал [-1, 1]. Каэфіцыент карэляцыі для нашых колькасны

прыклад Cos (V1, V2) = Cos (180 °) = -1, як і чакалася ад малюнку вышэй.

Адлегласць паміж двума наборамі дадзеных пункту V1 і V2 таксама інтэрнэт-прадукта:

$$| V1 - V2 | = (V1-V2) \cdot (V1-V2) = | V1 | 2+ | V2 | 2 - 2 | V1 | ' | V2 | = п [Var (V1)+ VarV2 - 2Cov (V1, V2)]$$

Зараз, пабудаваць матрыцу, слупкамі якой з'яўляюцца каардынаты двух вектараў V1 і V2 адпаведна. Памнажаючы Транспанаванне гэтай матрыцы само па сабе дае новы сіметрычнай матрыцы, якая змяшчае п раз дысперсіі V1 і V2 дысперсіі ў якасці сваёй асноўнай дыяганальных элементаў (напрыклад, 8, 2), і я раз Coy (V1, V2), а яго дыяганаль ад элементаў (напрыклад, 4).

Магчыма, вы захочаце выкарыстоўваць міліметровую паперу і калькулятар, каб праверыць вынікі гэтых лікавых прыкладах, а таксама выканаць некаторыя дадатковыя лікавых эксперыментаў для больш глыбокага разумення канцэпцыі.

Дадатковая літаратура:
Викенс Т., *геаметрыя шматмернага статыстычнага аналізу*, Erlbaum Pub, 1995..

Што такое сярэдняе геаметрычнае

Сярэдняе геаметрычнае п неадмоўнае лікавыя значэння з'яўляецца корань з творы значэнняў п. Назоўнік каэфіцыента карэляцыі Пірсана з'яўляецца сярэднім геаметрычным двух дысперсій. Гэта карысна для асерадненні "прадукт момант" каштоўнасцяў.

Выкажам здагадку, у вас ёсць два пазітыўных дадзеных кропак x і y, то сярэдняе геаметрычнае гэтых лікаў з'яўляецца лік (г) такая, што x/г = ў/б, і сярэдняе арыфметычнае (а) такое, што x - a = - y.

Геаметрычныя сродкі шырока выкарыстоўваюцца ў ЗША Бюро працоўнай статыстыкі ["Geomeans", як іх называюць] пры разліку індэкса спажывецкіх цэн у ЗША. Geomeans таксама выкарыстоўваюцца ў азначніках коштаў. Выкарыстанне статыстычных сярэдняе геаметрычнае для індэксаў, такіх як ідэальны індэкс Фішэра.

Калі некаторыя значэння з'яўляюцца вельмі вялікімі па велічыні і іншыя малыя, то сярэдняе геаметрычнае з'яўляецца лепш сярэдняга. У геаметрычнай прагрэсіі, найбольш значных сярэдняе геаметрычнае сярэдняе. Сярэдняе арыфметычнае вельмі зрушаныя ў бок большай колькасці ў серыю.

У якасці прыкладу, выкажам здагадку, што продажы некаторага павелічэння пункт да 110% у першы год і да 150%, што на другі год. Для прастаты выкажам здагадку, што вы прадалі 100 адзінак на пачатковым этапе. Тады лік прададзеных у першы год складае 110, а колькасць прададзеных ў другім складае 150% x 110 = 165. Сярэдняе арыфметычнае 110% і 150% 130%, так што мы няправільна лічым, што колькасць прададзеных у першы год складае 130, а лік у другі год 169. Сярэдняе геаметрычнае 110% і 150% г = (1.65) 1/2так, што мы правільна ацаніць, што мы прадаем 100 (г) 2 = 165 пунктаў на працягу другога года.

У якасці яшчэ аднаго падобнага прыкладу, калі паявы фонд расце на 50% у год і на 50% у наступным годзе, і вы трымаеце прылада працягу абодвух гадоў, вы страцілі грошы ў канцы. За кожны даляр, вы пачалі з таго, зараз вы атрымалі 75с. Такім чынам, прадукцыйнасць адрозніваецца ад атрымання (50% -50%)/2 (0%). Гэта так жа, як змяняецца з множнікам (1.5 x 0.5) ½ = 0,866 у год. У мультыплікатыўны працэс, адно значэнне, якое можа быць заменены для кожнага набору значэнняў, каб даць той жа "агульны эфект" з'яўляецца сярэднім геаметрычным, а не сярэдняе арыфметычнае. Як правіла, грошы мультыплікатыўны ("ён бярэ грошы, каб зарабляць грошы"), фінансавыя дадзеныя, часта лепш спалучаюцца такім чынам.

Як прыклад аналізу апытання, даць прыклад людзей, спіс, скажам, 10, злачынстваў у дыяпазоне сур'ёзнасцю:

Крадзеж... Напад... Падпал.. Згвалтаванне... Забойства

Папытаеце кожнага рэспандэнта даваць якія-небудзь лікавыя значэння яны адчуваюць, што любое злачынства, у спісе (напрыклад, хтосьці можа прыняць рашэнне аб выкліку падпалу 100). Затым папытаеце іх ацаніць кожнае злачынства ў спісе на шкале адносін. Калі згвалтаванне адказчык думка была ў пяць разоў так дрэнна, як падпал, тое значэнне 500 будзе прызначаны, крадзяжы квартале дрэнна, 25. Выкажам здагадку, што мы хацелі, каб "сярэдні" рэйтынг сярод рэспандэнтаў дадзенай кожнага злачынства. З рэспандэнтаў выкарыстоўваюць свае ўласныя базавую значэнне, сярэдняе арыфметычнае было б бескарысна: людзі, якія выкарыстоўвалі вялікая колькасць іх базавага значэння будзе "балота" тыя, хто абраў невялікіх колькасцях. Тым не менш, сярэдняе геаметрычнае - корань з творы рэйтынг за кожнае злачынства рэспандэнтаў га - дае роўныя вагі для ўсіх адказаў. Я выкарыстаў гэта ў класе практыкаванні і гэта працуе выдатна.

Вельмі часта добры для ўваходу-пераўтварэння гэтых дадзеных да рэгрэсіі, Дысперсійны аналіз, і г.д. Гэтыя статыстычныя метады даюць заключэнне аб сярэдняй арыфметычнай (якая цесна звязана з метадам найменшых квадратаў памылкі вымярэння), аднак сярэдняе арыфметычнае лог-пераўтвораных дадзеных гэта часопіс сярэдняе геаметрычнае дадзеных. Так, напрыклад, пры выпрабаванні на ўваходзе-пераўтвораных дадзеных на самай справе з'яўляецца тэст для размяшчэння сярэдняе геаметрычнае.

Дадатковая літаратура:
Лэнглі Р. *Практычны Статыстыка проста патлумачыў*, 1970, Dover Press.

Што такое Цэнтральная лімітавая тэарэма?

Для практычных мэт, галоўная ідэя цэнтральнай лімітавай тэарэмы (ЦПТ) у тым, што сярэдні ўзор назірання ўзятыя з некаторых насельніцтва любой формы размеркавання прыблізна размяркоўваюцца нармальнае размеркаванне, калі выконваюцца пэўныя ўмовы. У тэарэтычнай статыстыцы існуе некалькі варыянтаў цэнтральнай лімітавай тэарэмы ў залежнасці ад таго, наколькі гэтыя ўмовы пазначаныя. Яны звязаныя з тыпамі дапушчэнняў аб размеркаванні бацькоўскага насельніцтва (колькасць насельніцтва, з якіх выбарка) і ўласна працэдуру адбору.

Адзін з самых простых версій тэарэма сцвярджае, што калі выпадковая выбарка аб'ёму п (скажам, п > 30) ад бясконцага канчатковай сукупнасці стандартнае адхіленне, то стандартызаваныя ўзоры азначае сыходзіцца да стандартнага нармальнаму размеркаванні ці, эквівалентна, ўзору азначае, набліжаецца да нармальнага размеркаванні з сярэднім значэннем роўным сярэдняму насельніцтва і стандартнае адхіленне роўна стандартнае адхіленне насельніцтва падзелена на квадратны корань з ўзору N памеру. Пры ўжыванні цэнтральнай лімітавай тэарэмы для практычных задач статыстычнага вываду, аднак, статыстыкі больш зацікаўлены ў тым, як цесна прыкладныя размеркаванне выбарачнага сярэдняга наступным нармальнага размеркавання для канчатковых памераў выбаркі, чым гранічнае размеркаванне сабе. Дастаткова добра ўзгадняецца з нармальным размеркаваннем статыстыкі дазваляе выкарыстоўваць звычайныя тэорыя робіць высновы аб насельніцтве параметраў (такіх, як сярэдняя),

выкарыстоўваючы выбарачнае сярэдняе, незалежна ад канкрэтнага віду бацькоўскай папуляцыі.

Добра вядома, што ўсе бацькі насельніцтва, стандартызаваная пераменная будзе мець размеркаванне з сярэднім 0 і стандартным адхіленнем 1 па выпадковай выбарцы. Больш за тое, калі бацька насельніцтва нармальна, то распаўсюджваецца гэтак жа, як стандартная нармальная пераменная для любога натуральнага п лік. Цэнтральная лімітавая тэарэма сцвярджае, выдатны вынік, што, нават калі бацька насельніцтва не з'яўляецца нармальным, стандартызаваных зменных прыблізна нармальным, калі выбарка досыць вялікая (скажам, > 30). Як правіла, не дазваляюць сцвярджаць ўмовы, пры якіх набліжэнне дае цэнтральная лімітавая тэарэма працуе, а што памер выбаркі неабходна да набліжэння становіцца досыць добрым. У якасці агульнай рэкамендацыі, статыстыкаў выкарыстоўвалі рэцэпт, што калі бацькі размеркавання сіметрычна і адносна кароткім хвастом, то выбарачнае сярэдняе дасягае прыблізна нармалёвасці для невялікіх узораў, чым калі б бацькі насельніцтва скажонае або доўгі хвост.

Пра электроннай павінна вывучаць паводзіны сярэдніх узораў розных памераў, атрыманыя з розных матчыных насельніцтва. Вывучэнне выбаркі размеркаванне выбарачных сярэдніх вылічаецца з узораў розных памераў, атрыманыя з розных размеркаванняў, дазваляюць атрымаць некаторы ўяўленне аб паводзінах ўзору азначае, у тых канкрэтных умовах, а таксама вывучыць дзеянні кіруючых прынцыпаў, згаданых вышэй, для выкарыстання Цэнтральная лімітавая тэарэма ў практыку.

Пры пэўных умовах у вялікіх выбарках, выбарачнае размеркаванне выбарачнага сярэдняга можна аппроксимировать нармальным размеркаваннем. Памер выбаркі, неабходных для набліжэння да дастатковым моцна залежыць ад формы бацькоўскага размеркавання. Сіметрыі (або яе адсутнасць) мае асаблівае значэнне. Для сіметрычнага размеркавання бацькоў, нават калі моцна адрозніваецца ад формы нармальнага размеркавання, адэкватнага набліжэння могуць быць атрыманы пры малых выбарак (напрыклад, 10 або 12 для раўнамернага размеркавання). Для сіметрычных кароткіх хвастамі бацькоў, выбарачнае сярэдняе дасягае прыблізна нармалёвасці для невялікіх узораў, чым калі б бацькі насельніцтва скажонае і доўгі хвост. У некаторых крайніх выпадках (напрыклад, биномиальные с) ўзоры памерамі нашмат перавышае тыповыя кіруючыя прынцыпы (напрыклад, 30), неабходных для адэкватнага набліжэння. Для некаторых дыстрыбутываў без першага і другога момантаў (напрыклад, Коши), цэнтральная лімітавая тэарэма не мае месца.

Што такое выбарачнае размеркаванне?

Асноўная ідэя статыстычнага вываду, узяць выпадковую выбарку з генеральнай сукупнасці, а затым выкарыстоўваць гэтую інфармацыю ад ўзору, каб зрабіць высновы аб канкрэтных характарыстыках насельніцтва, такія як сярэдняе (мера цэнтральнай тэндэнцыі), стандартнае адхіленне (мера роскіду) або долі ў генеральнай сукупнасці, якія маюць пэўныя характарыстыкі. Выбарка эканоміць грошы, час і намаганні. Акрамя таго, прыклад, у некаторых выпадках прадастаўляць столькі ж ці большай дакладнасцю, чым адпаведныя даследаванні, якія будуць спрабаваць даследаваць ўсё насельніцтва, пільны збор дадзеных ад ўзору будуць часта даюць больш інфармацыі, чым менш дбайнага вывучэння, які спрабуе выглядаць ва ўсім.

Мы будзем вывучаць паводзіны сярэдніх выбарачных значэнняў з іншай названай групы насельніцтва. Таму што прыклад разглядае толькі частка насельніцтва, выбарачнае сярэдняе не будзе ў дакладнасці роўная адпаведнай сярэдняй часткі насельніцтва. Такім чынам, важным фактарам для тых, хто плануе і інтэрпрэтацыі вынікаў выбаркі, з'яўляецца ступень, у якой ўзор ацэнак, такія як выбарачнае сярэдняе, пагодзяцца з адпаведнымі характарыстыка насельніцтва.

На практыцы толькі адзін прыклад, як правіла, прымаюцца (у некаторых выпадках невялікая `` пілот" ўзор выкарыстоўваецца для праверкі дадзеных механізмаў збору і атрымання папярэдняй інфармацыі для планавання асноўнай схемай выбаркі). Тым не менш, у мэтах разумення, у якой ступені прыклад сродкі будуць згодныя з адпаведным насельніцтвам маю на ўвазе, што гэта карысна разгледзець, што адбудзецца, калі 10 ці 50, ці 100 асобных даследаванняў, адбору пробаў, таго ж тыпу, былі праведзеныя. Наколькі паслядоўна будуць вынікі будуць праз гэтыя розныя даследаванні? Калі б мы маглі бачыць, што вынікі кожнага з узораў быў бы амаль тое ж самае (і амаль правільна!), То мы хацелі б быць упэўненыя ў адным узоры, што на самой справе будзе выкарыстоўвацца. З іншага боку, бачачы, што адказы ад паўтаральных узораў былі занадта пераменная неабходная дакладнасць мяркуе, што іншага плана выбаркі (магчыма, з вялікім памерам выбаркі) павінны быць выкарыстаны.

Выбарачнае размеркаванне выкарыстоўваецца для апісання размеркавання вынікаў, якія можна было б назіраць з рэплікацыі канкрэтнага плана выбаркі.

Ведайце, што для ацэнкі сродкаў годнасці (каб даць значэнне).

Ведайце, што ацэнкі вылічаюцца з аднаго ўзору будзе адрозніваца ад ацэнкі, якая будзе вылічана з іншага ўзору.

Зразумейце, што ацэнкі, як чакаецца, адрозніваюцца ад папуляцыйных характарыстык (параметраў), што мы спрабуем ацаніць, але, што ўласцівасці выбарачных размеркаванняў дазваляюць нам ацаніць, імавернасным, як яны будуць адрозніваца.

Зразумейце, што розныя маюць розныя размеркавання выбаркі з размеркаваннем форму ў залежнасці ад (а) канкрэтнай статыстыкі, (б) памер выбаркі, і (у) бацька размеркавання.

Разуменне ўзаемасувязі паміж памерам выбаркі і размеркавання выбарачных ацэнак.

Зразумейце, што адрозненні ў размеркаванні выбаркі можа быць паменшаны за кошт павелічэння памеру выбаркі.

Адзначым, што ў вялікіх выбарках, многія дыстрыбутывы выбаркі можна наблізіць з нармальным размеркаваннем.

Выкід выдаленне

Выкіды некалькі заўваг, якія не вельмі добра абсталяваны ў "лепшых" даступныя мадэлі. На практыцы любое заўвага са стандартнымі рэшткавым больш за 2,5 па абсалютнай велічыні з'яўляецца кандыдатам на выкід. У такім выпадку трэба спачатку даследаваць крыніца дадзеных, калі ёсць сумневы ў дакладнасці і пэўнасці назіранняў, то яна павінна быць выдаленая і мадэлі павінны быць пераабсталяваны.

Надзейныя статыстычныя метады неабходныя, каб справіцца з любой незаўважанымі выкідаў, у адваротным выпадку вынік будзе ўводзіць у зман. Напрыклад, звычайны пакрокавай рэгрэсіі часта выкарыстоўваецца для выбару адпаведнага падмноства незалежных зменных для выкарыстання ў мадэлі, аднак, ён можа быць прызнаны несапраўдным, нават наяўнасць некалькіх выкідаў.

У сувязі з патэнцыйна вялікай розніцы, выкіды могуць быць вынікам адбору. Гэта зусім правільна, каб такія назірання, што законна належыць даследчай групы па азначэнні.Логнормально размеркаваных дадзеных (напрыклад, курс міжнароднай), напрыклад, будзе часта праяўляюць такія значэння.

Такім чынам, вы павінны быць вельмі ўважлівыя і асцярожныя: перш чым аб'явіць заўвагу "выкід", высветліць, чаму і як такое назіранне было. Гэта можа быць нават памылкі на этапе ўводу дадзеных.

Па-першае, пабудаваць BoxPlot дадзеных. Форма Q1, Q2 і Q3 кропак, якія дзеляць узораў на чатыры аднолькавых па памеры групы. (Q2 = сярэдні) Няхай МКР = Q3 - Q1. Выкіды вызначаюцца як тыя кропкі, па-за межамі значэння Q3+ да * IQR і Q1-да * МКР. У большасці выпадку ўстанаўліваецца да = 1,5.

Іншы альтэрнатывай з'яўляецца наступны алгарытм

а) Вылічыць з цэлай ўзору.
б) Вызначыць набор межах ад сярэдняй: сярэдняя+ да з, сярэдняя - да з.. сігма (Дазволіць карыстачу ўвайсьці да тыповым значэннем для да-2)
в) Выдаліце ўсе ўзоры значэння за межы.

Цяпер ітэрацыі N раз па алгарытму, кожны раз замяняючы выбарачнай сукупнасці з памяншэннем узораў пасля нанясення крок (с).

Звычайна нам трэба для выканання ітэрацыі гэтага алгарытму ў 4 разы.

Як згадвалася раней, агульны "стандартны" любое назіранне за падзеннем 1.5 (верагоднае адхіленне), т. е. (1,5 IQRs) вагаецца вышэй 3. Кватэр ці ніжэй 1. Кватэр. У наступнай праграме SPSS, дапаможа вам у вызначэнні выкідаў.

```
$ SPSS/OUTPUT = LIER.OUT    назвай "Вызначэнне, калі выкіды існуе"    DATA LIST Free File = 'A'/X1    VAR лэйблам          "Уваходныя д
```

Выкід выяўлення ў адным месцы насельніцтва была падрабязна разгледжана ў літаратуры. Нярэдка, аднак, можна сцвярджаць, што выяўленыя выкіды на самай справе не выкіды, а ўтвараюць 2. Насельніцтва. Калі гэта так, то кластарны падыход павінны быць прынятыя. Гэта будзе актыўны вобласці даследаванняў для вывучэння гэтай праблемы, як выкіды могуць паўстаць і вызначыць, калі кластарны падыход павінны быць прынятыя.

Дадатковая літаратура:
Хокінс Д. *выяўленні выкідаў*., Чэпмен і Хол, 1980
Ротамстед В. В. Барнета, Т. Люіс, *Выкіды ў статыстычных дадзеных*, М., 1994.

Найменшых квадратаў мадэлі

Многія праблемы звязаны з аналізам даных, якія апісваюць зменныя звязаныя паміж сабой. Самы просты з усіх мадэляў, якія апісваюць ўзаемасувязь паміж двума зменнымі з'яўляецца лінейнай, ці прамой лініі, мадэлі. Найпросты спосаб устаноўкі лінейнай мадэлі з'яўляецца `` вачэй" мяч лінію праз дадзеныя па сюжэце, але больш хупавы і традыцыйны метад з'яўляецца тое, што найменшых квадратаў, які знаходзіць лінію мінімізацыі сумы адлегласцяў паміж назіраецца кропкі і абсталяваны лініі.

Зразумейце, што ўстаноўкі `` лепшы" лінію на вока цяжка, асабліва, калі ёсць шмат рэшткавым зменлівасці дадзеных.

Ведайце, што існуе просты сувязі паміж лікавыя каэфіцыенты ў ўраўненні рэгрэсіі і нахілу і адрэзак лініі рэгрэсіі.

Ведайце, што адна статыстыка рэзюмэ, як каэфіцыент карэляцыі ці не распавесці ўсю гісторыю. Кропкавая дыяграма з'яўляецца істотным дадаткам да вывучэння сувязі паміж гэтымі дзвюма зменнымі.

Ведайце, што мадэль праверкі з'яўляецца неад'емнай часткай працэсу статыстычнага мадэлявання. У рэшце рэшт, высновы, заснаваныя на мадэлях, якія не належным чынам апісаць назіраныя набор дадзеных будзе несапраўдным.

Ведаць ўплыў парушэнні здагадкі Рэгрэсійная мадэлі (напрыклад, умовы) і магчымыя рашэнні на аснове аналізу рэшткаў.

Найменш Медыяна квадратаў мадэлі

Стандартныя метады найменшых квадратаў для ацэнкі лінейнай мадэлі не з'яўляюцца надзейнымі у тым сэнсе, што выкіды або забруджаныя дадзеныя могуць моцна ўплываць на ацэнкі. Надзейная тэхніка, якая абараняе ад забруджвання як мінімум сярэдняе квадратаў (LMS) ацэнкі. Пашырэнне LMS ацэнку абагульненай лінейнай мадэлі, што прыводзіць да найменш сярэднім адхіленне (LMD) ацэнкі.

Што такое дастаткова?

Дастатковай ацэнкі на аснове статыстыкі змяшчае ўсю інфармацыю, якая прысутнічае ў зыходных дадзеных. Напрыклад, сума дадзеных дастаткова ацаніць сярэднюю частку насельніцтва. Вы не павінны ведаць дадзеныя перад сабой. Гэта эканоміць шмат грошай, калі дадзеныя павінны быць перададзены па сетцы электрасувязі. Прасцей кажучы, адправіць агульнага і памер выбаркі.

Дастатковай статыстыкай т для параметру Q з'яўляецца функцыяй x1 выбаркі дадзеных,..., хп, які змяшчае ўсю інфармацыю ў узоры адносна параметру ц. Больш фармальна, дастатковасці вызначаецца ў тэрмінах функцыі праўдападобнасьці для д. Для дастатковага т статыстыка, верагоднасць таго, L (x1,..., x | д) можа быць запісана ў выглядзе

г (т | д) * да (x1,..., хп)

Паколькі 2. Складнік не залежыць ад д, т завецца дастатковай статыстыкай для ц.

Іншы спосаб пастаноўкі гэтага для звычайных праблем з'яўляецца тое, што можна было б пабудаваць выпадковы працэс, пачынаючы з дастатковай статыстыкай, якая будзе мець дакладна такое ж размеркаванне, як поўны ўзор для ўсіх дзяржаў характар.

Каб праілюстраваць гэта, давайце заўвагі незалежных выпрабаванняў Бярнулі з аднолькавай верагоднасцю поспеху. Выкажам здагадку, што існуе п выпрабаванняў, і чалавек заўважае, якое назірання поспехаў, і чалавек В толькі пазнае, што колькасць поспехаў. Тады, калі В змяшчае гэтыя поспехі ў выпадковых кропках без рэплікацыі, верагоднасць таго, што В цяпер атрымаць любы набор поспехі сапраўды гэтак жа, як верагоднасць таго, што ўбачыце, што набор, незалежна ад таго, якая праўдзівая верагоднасць поспеху бывае.

Вы павінны глядзець на Ваш Scattergrams!

Даведайцеся, што дадзены набор дадзеных лініі рэгрэсіі з'яўляецца унікальным. Тым не менш, зваротны гэта сцвярджэнне не адпавядае рэчаіснасці. Наступны цікавы прыклад з, Д. Мур (1997), кніга, стар 349:

Набор дадзеных: x 10 8 13 9 11 14 г 8,04 6,95 7,58 8,81 8,33 9,96 x 6 4 12 7 5 г 7,24 4,26 10,84 4,82 5,68 Набор дадзеных в:

Усе тры падыходу маюць тыя ж карэляцыі і рэгрэсіі. Важным з'яўляецца маральным ***погляд на вашы scattergrams***.

Як стварыць лічбавай прыклад, дзе дзве дыяграмы рассейвання ясна паказваюць розныя адносіны (моцныя бакі), але даюць *ж* ковариация? Выканайце наступныя дзеянні:

1. Прадукцыя двух набораў (X, Y) значэнні, якія маюць іншае суадносіны ў,
2. Разлічыць двух ковариации, скажам, C1 і C2,
3. Выкажам здагадку, вы хочаце, каб зрабіць C2 роўная C1. Затым неабходна памножыць C2 (C1/C2),
4. Так як C = R x. S ў, трэба два нумары (адзін з іх можа быць 1), і б такія, што AB = (C1/C2),
5. Памножыць ўсе значэння X у камплекце 2, і ўсе значэння Y па б: для новых зменных, C = rabS x S. ў = C2 (C1/C2) = C1..

Цікавы колькасны прыклад, які паказвае, дыяграмы рассейвання двух аднолькавых, але з рознымі ковариации заключаецца ў наступным: разгледзім набор дадзеных (X, Y) значэнні, з ковариационной C1. Хай цяпер V = 2X, і W = 3г. Ковариация V і W будзе 2 (3) = 6 разоў C1, але суадносіны паміж V і W гэтак жа, як сувязь паміж X і Y.

Power of a Test

Significance tests are based on certain assumptions: The data have to be random samples out of a well defined basic population and one has to assume that some variables follow a certain distribution - in most cases the normal distribution is assumed.

Power of a test is the probability of correctly rejecting a false null hypothesis. This probability is one minus the probability of making a Type II error (b). Recall also that we choose the probability of making a Type I error when we set a and that if we decrease the probability of making a Type I error we increase the probability of making a Type II error.

Power and Alpha:

Therefore, the probability of correctly retaining a true null has the same relationship to Type I errors as the probability of correctly rejecting an untrue null does to Type II error. Yet, as I mentioned if we decrease the odds of making one type of error we increase the odds of making the other type of error. What is the relationship between Type I and Type II errors?

Power and the True Difference between Population Means: Anytime we test whether a sample differs from a population or whether two sample come from 2 separate populations, there is the assumption that each of the populations we are comparing has it's own mean and standard deviation (even if we do not know it). The distance between the two population means will affect the power of our test.

Power as a Function of Sample Size and Variance: You should notice that what really made the difference in the size of b is how much overlap there is in the two distributions. When the means are close together the two distributions overlap a great deal compared to when the means are farther apart. Thus, anything that effects the extent the two distributions share common values will increase b(the likelihood of making a Type II error).

Sample size has an indirect effect on power because it affects the measure of variance we use to calculate the t-test statistic. Since we are calculating the power of a test that involves the comparison of sample means, we will be more interested in the standard error (the average difference in sample values) than standard deviation or variance by itself. Thus, sample size is of interest because it modifies our estimate of the standard deviation. When n is large we will have a lower standard error than when n is small. In turn, when N is large well have a smaller b region than when n is small.

Pilot Studies: When the needed estimates for sample size calculation is not available from existing database, a pilot study is needed for adequate estimation with a given precision.

Further Readings:

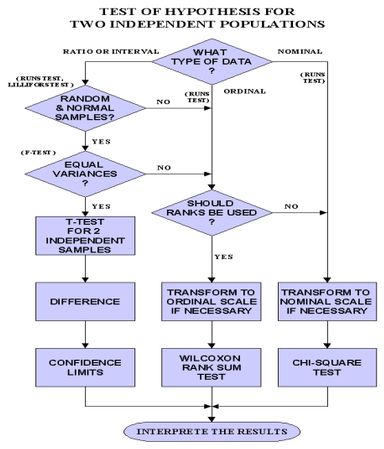
Cohen J., *Statistical Power Analysis for the Behavioral Sciences*, L. Erlbaum Associates, 1988.
Kraemer H., and S. Thiemann, *How Many Subjects?* Provides basic sample size tables, explanations, and power analysis.
Murphy K., and B. Myors, *Statistical Power Analysis*, L. Erlbaum Associates, 1998. Provides a simple and general sample size determination for hypothesis tests.

ANOVA: Analysis of Variance

The tests we have learned up to this point allow us to test hypotheses that examine the difference between only two means. Analysis of Variance or ANOVA will allow us to test the difference between 2 or more means. ANOVA does this by examining the ratio of variability between two conditions and variability within each condition. For example, say we give a drug that we believe will improve memory to a group of people and give a placebo to another group of people. We might measure memory performance by the number of words recalled from a list we ask everyone to memorize. A t-test would compare the likelihood of observing the difference in the mean number of words recalled for each group. An ANOVA test, on the other hand, would compare the variability that we observe between the two conditions to the variability observed within each condition. Recall that we measure variability as the sum of the difference of each score from the mean. When we actually calculate an ANOVA we will use a short-cut formula.

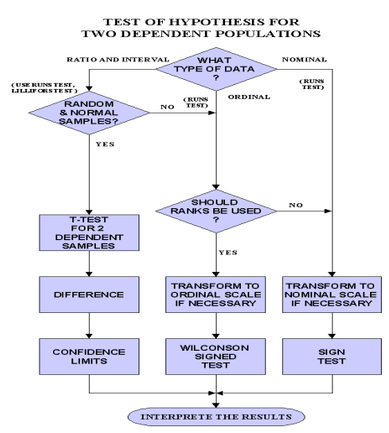
Thus, when the variability that we predict (between the two groups) is much greater than the variability we don't predict (within each group) then we will conclude that our treatments produce different results.

Levene's Test: Suppose that the sample data does not support the homogeneity of variance assumption, however, there is a good reason that the variations in the population are almost the same, then in such a situation you may like to use the Levene's modified test: In each group first compute the absolute deviation of the individual values from the median in that group. Apply the usual one way ANOVA on the set of deviation values and then interpret the results.



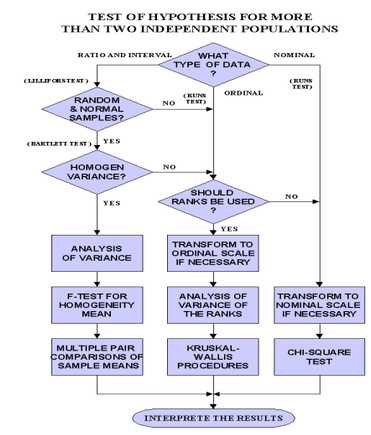
The Procedure for Two Populations Independent Means Test
Click on the image to enlarge it and THEN print it

You may use the following JavaScript to Test of Hypothesis for Two Populations



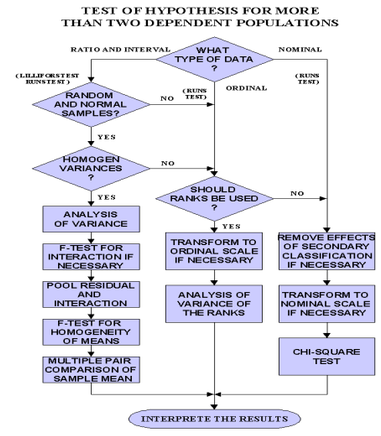
The Procedure for Two Dependent Means Test
Click on the image to enlarge it and THEN print it

You may use the following JavaScript to Two Dependent Populations Testing.



The Procedure for More Than Two Independent Means Test
Click on the image to enlarge it and THEN print it

You may use the following JavaScript to Three Means Comparison, Equality of Several Means' Test



The Procedure for More Than Two Dependent Populations Test
Click on the image to enlarge it and THEN print it

You may use the following JavaScript to Three Dependent Means Comparison.

Orthogonal Contrasts of Means in ANOVA

In repeated measurement of the analysis of variance when the null hypothesis is rejected, we might be interested in multiple comparisons of means by the combinations of means, this is known as the orthogonal contrasting the means. A contrast of the means is said to be orthogonal if the weighting means sum to zero. For example, the contrast of
(mean1+ mean2)/2 - mean3 is orthogonal. Therefore, to determine if two different contrasts of means from the same experiment are orthogonal, add the

product of the weights to see if they sum to zero. If they do not sum to zero, then the two contrasts are not orthogonal and only one of them could be tested. The orthogonal contrasting allows us to compare each mean against all of the other means. There are several effective methods of orthogonal contrasting for applications in testing, constructing confidence intervals, and the partial F-test as the post-analysis statistical activities of the usual ANOVA.

Further Readings:
Kachigan S., *Statistical Analysis: An Interdisciplinary Introduction to Univariate & Multivariate Methods*, Radius Press, 1986.
Kachigan S., *Multivariate Statistical Analysis: A Conceptual Introduction*, Radius Press, 1991.

The Six-Sigma Quality

The total approach to quality is essential for competing in world markets. The ability of a firm to give customers what they want at the lowest total cost gives a company an advantage over its competitors.

Sigma is a Greek symbol, which is used in statistics to represent standard deviation of a population. When a large enough random sample data are close to their mean (i.e., the average), then the population has a small deviation. If the data varies significantly from the mean, the data has a large deviation. In quality control measurement terms, you want to see that the sample is as close as possible to the mean and that the mean meets or exceed specifications. A large sigma means that there is a large amount of variation within the data. A lower sigma value corresponds to a small variation, and therefore a controlled process with a good quality.

The Six-Sigma means a measure of quality that strives for near perfection. Six-Sigma is a data-driven approach and methodology for eliminating defects to achieve six sigmas between lower and upper specification limits. Accordingly, to achieve Six-Sigma, e.g., in a manufacturing process it must not produce more than 3.4 defects per million opportunities. Therefore, a Six-Sigma defect is defined for not meeting the customer's specifications. A Six-Sigma opportunity is then the total quantity of chances for a defect.

Six-Sigma is a statistical measure expressing how close a product comes to its quality goal. One sigma means only 68% of products are acceptable; three sigma means 99.7% are acceptable. Six-Sigma is 99.9997% perfect or 3.4 defects per million parts or opportunities. The natural spread is 6 times the sample standard deviation. The natural spread is centered on the sample mean, and all weights in the sample fall within the natural spread, meaning the process will produce relatively few out-of-specification products. Six-Sigma does not necessarily imply 3 defective units per million made; it also signifies 3 defects per million opportunities when used to describe a process. Some products may have tens of thousands of opportunities for defects per finished item, so the proportion of defective opportunities may actually be quite large.

Six-Sigma Quality is a fundamental approach to delivering very high levels of customer satisfaction through disciplined use of data and statistical analysis for maximizing and sustaining business success. What that means is that all business decisions are made based on statistical analysis, not instinct or past history. Using the Six-Sigma approach will result in a significant, quantifiable improvement.

Is it truly necessary to go for zero defects? Why isn't 99.9% (about 4.6 sigma) defect-free good enough? Here are some examples of what life would be like if 99.9% were good enough:

- 1 hour of unsafe drinking water every month
- 2 long or short landings at every American cities airport each day
- 400 letters per hour which never arrive at their destination
- 3,000 newborns accidentally falling from the hands of nurses or doctors each year
- 4,000 incorrect drug prescriptions per year
- 22,000 checks deducted from the wrong bank account each hour

As you can see, sometimes 99.9% good just isn't good enough.

Here are some examples of what life would be still like at Six-Sigma, 99.9997% defect-free:

- 13 wrong drug prescriptions per year
- 10 newborns accidentally falling from the hands of nurses or doctors each year
- 1 lost article of mail per hour

Now we see why the quest for Six-Sigma quality is necessary.

Six-Sigma is the application of statistical methods to business processes to improve operating efficiencies. It provides companies with a series of interventions and statistical tools that can lead to breakthrough profitability and quantum gains in quality. Six-Sigma allows us to take a real world problem with many potential answers, and translate it to a math problem, which will have only one answer. We then convert that one mathematical solution back to a real world solution.

Six-Sigma goes beyond defect reduction to emphasize business process improvement in general, which includes total cost reduction, cycle-time improvement, increased customer satisfaction, and any other metric important to the customer and the company. An objective of Six-Sigma is to eliminate any waste in the organization's processes by creating a road map for changing data into knowledge, reducing the amount of stress companies experience when they are overwhelmed with day-to-day activities and proactively uncovering opportunities that impact the customer and the company itself.

The key to the Six-Sigma process is in eliminating defects. Organizations often waste time creating metrics that are not appropriate for the outputs being measured. Executives can get deceptive results if they force all projects to determine a one size fits all metric in order to compare the quality of products and services from various departments. From a managerial standpoint, having one universal tool seems beneficial; however, it is not always feasible. Below is an example of the deceptiveness of metrics.

In the airline industry, the US Air Traffic Control System Command Center measures companies on their rate of on time departure. This would obviously be a critical measurement to customers—the flying public. Whenever an airplane departs 15 minutes or more later than scheduled, that event is considered as a defect. Unfortunately, the government measures the airlines on whether the plane pulls away from the airport gate within 15 minutes of scheduled departure, not when it actually takes off. Airlines know this, so they pull away from the gate on time but let the plane sit on the runway as long as necessary before take off. The result to the customer is still a **late departure**. This defect metric is therefore not an accurate representation of the desires of the customers who are impacted by the process. If this were a good descriptive metric, airlines would be measured by the actual delay experienced by passengers.

This example shows the importance of having the right metrics for each process. The method above creates no incentive to reduce actual delays, so the customer (and ultimately the industry) still suffers. With a Six-Sigma business strategy, we want to see a picture that describes the true output of a process over time, along with additional metrics, to give an insight as to where the management has to focus its improvement efforts for the customer.

The Six Steps of Six-Sigma Loop Process: The process is identified by the following five major activities for each project:

1. Identify the product or service you provide—What do you do?
2. Identify your customer base, and determine what they care about—Who uses your products and services? What is really important to them?
3. Identify your needs—What do you need to do your work?
4. Define the process for doing your work—How do you do your work?
5. Eliminate wasted efforts—How can you do your work better?
6. Ensure continuous improvement by measuring, analyzing, and controlling the improved process—How perfectly are you doing your customer-focused work?

Often each step can create dozens of individual improvement projects and can last for several months. It is important to go back to each step from time to time in order to determine actual data maybe with improved measurement systems.

Once we know the answers to the above questions, we can begin to improve the process. The following case study will further explain the steps applied in Six-Sigma to Measure, Analyze, Improve, and Control a process to ensure customer satisfaction.

The Six Sigma General Process and Its Implementation: The Six-Sigma means a measure of quality that strives for near perfection. Six-Sigma is a data-driven approach and methodology for eliminating defects to achieve six-sigma's between lower and upper specification limits. Accordingly, to achieve Six-Sigma, e.g., in a manufacturing process it must not produce more than 3.4 defects per million opportunities. Therefore, a Six-Sigma defect is defined for not meeting the customer's specifications. A Six-Sigma opportunity is then the total quantity of chances for a defect. The implementation of the Six Sigma system starts normally with a few days workshop of the top level management of the organization.

Only if the advantages of Six Sigma can be clearly stated and supported of the entire Management, then it makes sense to determine together the first project surrounding field and the pilot project team.

The pilot project team members participate is a few days Six Sigma workshop to learn the system principals, the process, the tools and the methodology.

The project team meets to compiles main decisions and identifying key stakeholders in the pilot surrounding field. Within the next days the requirements of the stakeholders are collected for the main decision processes by face-to-face interviews.

By now, the workshop of the top management must be ready for the next step. The next step for the project team is to decide which and how the achievements should be measured and then begin with the data collection and analysis. Whenever the results are understood well then suggestions for improvement will be collected, analyzed, and prioritized based on the urgency and inter-dependencies.

As the main outcome, the project team members will determine which improvements should be realized first. In this phase it is important that rapid successes are obtained, in order to even the soil for other Six Sigma projects in the organization.

The activities must be carried out in parallel whenever possible by a network activity chart. The activity chart will become more and more realistic by a loop-process while spread the improvement throughout the organization. More and more processes will be included and employees are trained including Black Belts who are the six sigma masters, and the dependency of external advisors will be reduced.

The main objective of the Six-Sigma approach is the implementation of a measurement-based strategy that focuses on process improvement. The aim is variation reduction, which can be accomplished by Six-Sigma methodology.

The Six-Sigma is a business strategy aimed at the near-elimination of defects from every manufacturing, service and transactional process. The concept of Six-Sigma was introduced and popularized for reducing defect rate of manufactured electronic boards. Although the original goal of Six-Sigma was to focus on manufacturing process, today the marketing, purchasing, customer order, financial and health care processing functions also embarked on Six Sigma programs.

Motorola Inc.Case: Motorola is a role model for modern manufacturers. The maker of wireless communications products, semiconductors, and electronic equipment enjoys a stellar reputation for high-tech, high-quality products. There is a reason for this reputation. A participative-management process emphasizing employee involvement is a key factor in Motorola's quality push. In 1987, Motorola invested $44 million in employee training and education in a new quality program called Six-Sigma. Motorola measures its internal quality based on the number of defects in its products and processes. Motorola conceptualized Six-Sigma as a quality goal in the mid-1980. Their target was Six-Sigma quality, or 99.9997% defect free products—which is equivalent to 3.4 defects or less per 1 million parts. Quality is a competitive advantage because Motorola's reputation opens markets. When Motorola Inc. won the Malcolm Baldridge National Quality Award in 1988; it was in the early stages of a plan that, by 1992, would achieve Six-Sigma Quality. It is estimated that of $9.2 billion in 1989 sales, $480 million was saved as a result of Motorola's Six-Sigma program. Shortly thereafter, many US firms were following Motorola's lead.

Control Charts, and the CUSUM

Control charts for variables are called X- and R-charts. The X-charts is used to monitor the average variability and the R-chart is used to monitor the range of the variation.

Developing quality control charts for variables (X-Chart): The following steps are required for developing quality control charts for variables:

1. Decide what should be measured.
2. Determine the sample size.
3. Collect random sample and record the measurements/counts.
4. Calculate the average for each sample.
5. Calculate the overall average. This is the average of all the sample averages (X-double bar).
6. Determine the range for each sample.
7. Calculate the average range (R-bar).
8. Determine the upper control limit (UCL) and lower control limit (LCL) for the average and for the range.
9. Plot the chart.
10. Determine if the average and range values are in statistical control.
11. Take necessary action based on your interpretation of the charts.

Developing control charts for attributes (P-Chart): Control charts for attributes are called P-charts. The following steps are required to set up P-charts:

1. Determine what should be measured.
2. Determine the required sample size.
3. Collect sample data and record the data.
4. Calculate the average percent defective for the process (p).
5. Determine the control limits by determining the upper control limit (UCL) and the lower control limit (LCL) values for the chart.

6. Plot the data.
7. Determine if the percent defectives are within control.

Control charts are also used in industry to monitor processes that are far from Zero-Defect. However, among the powerful techniques is the counting of the cumulative conforming items between two nonconforming and its combined techniques based on cumulative sum and exponentially weighted moving average smoothing methods.

The general CUSUM is a statistical process control when measurements are multivariate. It is an effective tool in detecting a shift in the mean vector of the measurements, which is based on the cross-sectional antiranks of the measurements: At each time point, the measurements, after being appropriately transformed, are ordered and their antiranks are recorded. When the process is in-control under some mild regularity conditions the antirank vector at each time point has a given distribution, which changes to some other distribution when the process is out-of-control and the components of the mean vector of the process are not all the same. Therefore it detects shifts in all directions except the one that the components of the mean vector are all the same but not zero. This latter shift, however, can be easily detected by a univariate CUSUM.

Further Readings:
Breyfogle F., Implementing Six Sigma: Smarter Solutions Using Statistical Methods, Wiley, 1999.
del Castillo E., *Statistical Process and Adjustment Methods for Quality Control*, Wiley, 2002.
Juran J, and A. Godfreym, Juran's Quality Handbook, McGraw-Hill, 1999.
Xie M., T. Goh, and V. Kuralmani, *Statistical Models and Control Charts for High Quality Processes*, Kluwer, 2002.

Repeatability and Reproducibility

The term Repeatability refers to the equipment or instrument while Reproducibility refers to the equipment operator. Both Repeatability and Reproducibility involve statistical studies such as evaluation of statistical summaries, and comparison of the variances in repeat measurements, mostly for the industrial decision making problems. In these applications, for example the values indicated by the measuring devices vary from measurement to measurement. The main question is how much that built-in variation affects others activities, such as in-process measurements, quality checks, process improvement projects, etc.

Further Readings:
Barrentine L., *Concepts for R&R Studies*, ASQ Quality Press, 1991.
Wheeler D., and R. Lyday, *Evaluating the Measurement Process*, Statistical Process Control Press, 1990.

Statistical Instrument, Grab Sampling, and Passive Sampling Techniques

What is a statistical instrument? A statistical instrument is any process that aim at describing a phenomena by using any instrument or device, however the results may be used as a control tool. Examples of statistical instruments are questionnaire and surveys sampling.

What is grab sampling technique? The grab sampling technique is to take a relatively small sample over a very short period of time, the result obtained are usually instantaneous. However, the **Passive Sampling** is a technique where a sampling device is used for an extended time under similar conditions. Depending on the desirable statistical investigation, the Passive Sampling may be a useful alternative or even more appropriate than grab sampling. However, a passive sampling technique needs to be developed and tested in the field.

Distance Sampling

The term 'distance sampling' covers a range of methods for assessing wildlife abundance:

line transect sampling, in which the distances sampled are distances of detected objects (usually animals) from the line along which the observer travels

point transect sampling, in which the distances sampled are distances of detected objects (usually birds) from the point at which the observer stands

cue counting, in which the distances sampled are distances from a moving observer to each detected cue given by the objects of interest (usually whales)

trapping webs, in which the distances sampled are from the web center to trapped objects (usually invertebrates or small terrestrial vertebrates)

migration counts, in which the 'distances' sampled are actually times of detection during the migration of objects (usually whales) past a watch point

Many mark-recapture models have been developed over the past 40 years. Monitoring of biological populations is receiving increasing emphasis in many countries. Data from marked populations can be used for the estimation of survival probabilities, how these vary by age, sex and time, and how they correlate with external variables. Estimation of immigration and emigration rates, population size and the proportion of age classes that enter the breeding population are often important and difficult to estimate with precision for free-ranging populations. Estimation of the finite rate of population change and fitness are still more difficult to address in a rigorous manner.

Further Readings:
Buckland S., D. Anderson, K. Burnham, and J. Laake, *Distance Sampling: Estimating Abundance of Biological Populations*, Chapman and Hall, London, 1993.
Buckland S., D. Anderson, K. Burnham, J. Laake, D. Borchers, and L. Thomas, *Introduction to Distance Sampling*, Oxford University Press, 2001.

Data Mining and Knowledge Discovery

How to discover value in mountain of data? Data mining uses sophisticated statistical analysis and modelling techniques to uncover patterns and relationships hidden in organizational databases. Data mining and knowledge discovery aim at tools and techniques to process structured information from databases to data warehouses to data mining, and to knowledge discovery. Data warehouse applications have become business-critical. Data mining can compress even more value out of these huge repositories of information.

The continuing rapid growth of on-line data and the widespread use of databases necessitate the development of techniques for extracting useful knowledge and for facilitating database access. The challenge of extracting knowledge from data is of common interest to several fields, including statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing.

The data mining process involves identifying an appropriate data set to "mine" or sift through to discover data content relationships. Data mining tools include techniques like case-based reasoning, cluster analysis, data visualization, fuzzy query and analysis, and neural networks. Data mining sometimes resembles the traditional scientific method of identifying a hypothesis and then testing it using an appropriate data set. Sometimes however data mining is reminiscent of what happens when data has been collected and no significant results were found and hence an ad hoc, exploratory analysis is conducted to find a significant relationship.

Data mining is the process of extracting knowledge from data. The combination of fast computers, cheap storage, and better communication makes it easier by the day to tease useful information out of everything from supermarket buying patterns to credit histories. For clever marketers, that knowledge can be worth as much as the stuff real miners dig from the ground.

Data mining as an analytic process designed to explore large amounts of (typically business or market related) data in search for consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The process thus consists of three basic stages: exploration, model building or pattern definition, and validation/verification.

What distinguishes data mining from conventional statistical data analysis is that data mining is usually done for the purpose of "secondary analysis" aimed at finding unsuspected relationships unrelated to the purposes for which the data were originally collected.

Data warehousing as a process of organizing the storage of large, multivariate data sets in a way that facilitates the retrieval of information for analytic purposes.

Data mining is now a rather vague term, but the element that is common to most definitions is "predictive modeling with large data sets as used by big companies". Therefore, data mining is the extraction of hidden predictive information from large databases. It is a powerful new technology with great potential, for example, to help marketing managers "preemptively define the information market of tomorrow." Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools. Data mining answers business questions that traditionally were too time-consuming to resolve. Data mining tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining techniques can be implemented rapidly on existing software and hardware platforms across the large companies to enhance the value of existing resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client-server or parallel processing computers, data mining tools can analyze massive databases while a customer or analyst takes a coffee break, then deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

Knowledge discovery in databases aims at tearing down the last barrier in enterprises' information flow, the data analysis step. It is a label for an activity performed in a wide variety of application domains within the science and business communities, as well as for pleasure. The activity uses a large and heterogeneous data-set as a basis for synthesizing new and relevant knowledge. The knowledge is new because hidden relationships within the data are explicated, and/or data is combined with prior knowledge to elucidate a given problem. The term relevant is used to emphasize that knowledge discovery is a goal-driven process in which knowledge is constructed to facilitate the solution to a problem.

Knowledge discovery maybe viewed as a process containing many tasks. Some of these tasks are well understood, while others depend on human judgment in an implicit matter. Further, the process is characterized by heavy iterations between the tasks. This is very similar to many creative engineering process, e.g., the development of dynamic models. In this reference mechanistic, or first principles based, models are emphasized, and the tasks involved in model development are defined by:

1. Initialize data collection and problem formulation. The initial data are collected, and some more or less precise formulation of the modeling problem is developed.
2. Tools selection. The software tools to support modeling and allow simulation are selected.
3. Conceptual modeling. The system to be modeled, e.g., a chemical reactor, a power generator, or a marine vessel, is abstracted at first. The essential compartments and the dominant phenomena occurring are identified and documented for later reuse.
4. Model representation. A representation of the system model is generated. Often, equations are used; however, a graphical block diagram (or any other formalism) may alternatively be used, depending on the modeling tools selected above.
5. Computer implementation. The model representation is implemented using the means provided by the modeling system of the software employed. These may range from general programming languages to equation-based modeling languages or graphical block-oriented interfaces.
6. Verification. The model implementation is verified to really capture the intent of the modeler. No simulations for the actual problem to be solved are carried out for this purpose.
7. Initialization. Reasonable initial values are provided or computed, the numerical solution process is debugged.
8. Validation. The results of the simulation are validated against some reference, ideally against experimental data.
9. Documentation. The modeling process, the model, and the simulation results during validation and application of the model are documented.
10. Model application. The model is used in some model-based process engineering problem solving task.

For other model types, like neural network models where data-driven knowledge is utilized, the modeling process will be somewhat different. Some of the tasks like the conceptual modeling phase, will vanish.

Typical application areas for dynamic models are control, prediction, planning, and fault detection and diagnosis. A major deficiency of today's methods is the lack of ability to utilize a wide variety of knowledge. As an example, a black-box model structure has very limited abilities to utilize first principles knowledge on a problem. this has provided a basis for developing different hybrid schemes. Two hybrid schemes will highlight the discussion. First, it will be shown how a mechanistic model can be combined with a black-box model to represent a pH neutralization system efficiently. Second, the combination of continuous and discrete control inputs is considered, utilizing a two-tank example as case. Different approaches to handle this heterogeneous case are considered.

The hybrid approach may be viewed as a means to integrate different types of knowledge, i.e., being able to utilize a heterogeneous knowledge base to derive a model. Standard practice today is that almost any methods and software can treat large homogeneous data-sets. A typical example of a homogeneous data-set is time-series data from some system, e.g., temperature, pressure, and compositions measurements over some time frame provided by the instrumentation and control system of a chemical reactor. If textual information of a qualitative nature is provided by plant personnel, the data becomes heterogeneous.

The above discussion will form the basis for analyzing the interaction between knowledge discovery, and modeling and identification of dynamic models. In particular, we will be interested in identifying how concepts from knowledge discovery can enrich state-of-the- art within control, prediction, planning, and fault detection and diagnosis of dynamic systems.

Further Readings:

Marco D., *Building and Managing the Meta Data Repository: A Full Lifecycle Guide*, John Wiley, 2000.
Thuraisingham B., *Data Mining: Technologies, Techniques, Tools, and Trends*, CRC Press, 1998.
Westphal Ch., T. Blaxton, *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*, John Wiley, 1998.

Neural Networks Applications

Artificial neural networks provide a well-established, powerful tool to infer patterns from large databases. They have proven to be very useful to solve problems of interpolation, classification and prediction, and have been used in a vast number of business and financial applications.

The classical approaches are the feedforward neural networks, trained using back-propagation, which remain the most widespread and efficient technique to implement supervised learning. The main steps are: preprocess the data, the appropriate selection of variables, postprocessing of the results, and a final validation of the global strategy. Applications include data mining, and stock market predictions.

Further Readings:
Schurmann J., *Pattern Classification: A Unified View of Statistical and Neural Approaches*, John Wiley & Sons, 1996.

Bayes and Empirical Bayes Methods

Bayes and empirical Bayes (EB) methods structure combining information from similar components of information and produce efficient inferences for both individual components and shared model characteristics. Many complex applied investigations are ideal settings for this type of synthesis. For example, county-specific disease incidence rates can be unstable due to small populations or low rates. 'Borrowing information' from adjacent counties by partial pooling produces better estimates for each county, and Bayes/empirical Bayes methods structure the approach. Importantly, recent advances in computing and the consequent ability to evaluate complex models, have increase the popularity and applicability of Bayesian methods.

Bayes and EB methods can be implemented using modern Markov chain Monte Carlo(MCMC) computational methods. Properly structured Bayes and EB procedures typically have good frequentist and Bayesian performance, both in theory and in practice. This in turn motivates their use in advanced high-dimensional model settings (e.g., longitudinal data or spatio-temporal mapping models), where a Bayesian model implemented via MCMC often provides the only feasible approach that incorporates all relevant model features.

Further Readings:
Bernardo J., and A. Smith, *Bayesian Theory*, Wiley, 2000.
Carlin B., and T. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall, 1996.
Congdon P., *Bayesian Statistical Modelling*, Wiley, 2001.
Press S., and J. Tanur, *The Subjectivity of Scientists and the Bayesian Approach*, Wiley, 2001. Comparing and contrasting the reality of subjectivity in the work of history's great scientists and the modern Bayesian approach to statistical analysis.

Markovian & Memory Theory

According to the Memory (M) Theory, in modeling the memory events, events that depend on two or more past times, not just 1 as in Markov chains/processes, or none as in time-independent events, it is best to change ratios to differences (plus a constant - 1 is very nice, but other constants including 0 are often used). Ratios and products work best in Bayesian Markov chains and processes. Differences (subtraction) and sums work best in M Theory and M Events. These latter events range from viscoelastic materials through human memory to economic/financial/biological memory processes. Addition and subtraction has its own simplifications (e.g., geometric series sum exceptionally easily), and at an advanced level a special type of multiplication generalizes subtraction, namely the convolution product which is already widely recognized as being involved in memory (via Volterra integral and integral-differential equations, etc.). Volterra equations, by the way, are relatively easy to solve, and even numerical analysis/approximation software is just as available as main software if you know where to look (usually in the physical science/engineering software). The simplification due to convolution products is at least as great as the simplification involved in multiplicative ordinary multiplication, and allows advanced Fourier transform and Laplace transform methods to be used.

Memory Theory and time series share the additive property and inside a single term there can be multiplication, but like general regression methods this does not always mean that they are all using M Theory. One may use standard time series methods in the initial phase of modeling things, but instead proceed as follows using M Theory's Cross-Term Dimensional Analysis (CTDA). Suppose that you postulate a model y = af(x) - bg(z)+ ch(u) where f, g, h are some functions and x, z, u are what are usually referred to as independent variables. Notice the minus sign (-) to the left of b and the+ sign to the left of c and (implicitly) to the left of a, where a, b, c are positive constants. The variable y is usually referred to as a dependent variable. According to M Theory, not only do f, g, and h influence/cause y, but g influences/causes f and h at least to some extent. In fact, M Theory can formulate this in terms of probable influence as well as deterministic influence. All this generalizes to the case where the functions f, g, h depend on two or more variables, e.g., f(x, w), g(z, t, r), etc.

One can reverse this process. If one thinks that f influences g and h and y but that h and g only influence y and not f, then express the equation of y in the above form. If it works, one has found something that mainstream regression and time series may fail to detect. Of course, path analysis and Lisrel and partial least squares also claim to have ‘causal' abilities, but only in the standard regression sense of ‘freezing' so-called independent variables as ‘givens' and not in the M Theory sense which allows them to vary with y. In fact, Bayesian probability/statistics methods and M Theory methods use respectively ratios like y/x and differences like y - x+ 1 in their equations, and in the Bayesian model x is fixed but in the M Theory model x can vary. If one looks carefully, one will notice that the Bayesian model blows up at x = 0 (because division by 0 is impossible, visit the The Zero Saga page), but also near x = 0 since an artificially enormous increase is introduced - precisely near rare events. That is one of the reasons why M Theory is more successful for rare and/or highly influenced/influencing events, while Bayesian and mainstream methods work fairly well for frequent/common and/or low influence (even independent) and/or low dependence events.

Further Readings:
Kursunuglu B., S. Mintz, and A. Perlmutter, *Quantum Gravity, Generalized Theory of Gravitation, and Superstring Theory-Based Unification*, Kluwer Academic/Plenum, New York 2000.

Likelihood Methods

Direct	Inverse		Neyman-Pearson

In the Direct schools, one uses Pr(data | hypothesis), usually from some model-based sampling distribution, but one does not attempt to give the inverse probability, Pr(hypothesis | data), nor any other quantitative evaluation of hypotheses. The Inverse schools do associate numerical values with hypotheses, either probabilities (Bayesian schools) or something else (Fisher, Edwards, Shafer).

The decision-oriented methods treat statistics as a matter of action, rather than inference, and attempt to take utilities as well as probabilities into account in selecting actions; the inference-oriented methods treat inference as a goal apart from any action to be taken.

The "hybrid" row could be more properly labeled as "hypocritical"-- these methods talk some Decision talk but walk the Inference walk.

Fisher's fiducial method is included because it is so famous, but the modern consensus is that it lacks justification.

Now it is true, under certain assumptions, some distinct schools advocate highly similar calculations, and just talk about them or justify them differently. Some seem to think this is tiresome or impractical. One may disagree, for three reasons:

First, how one justifies calculations goes to the heart of what the calculations actually MEAN; second, it is easier to teach things that actually make sense (which is one reason that standard practice is hard to teach); and third, methods that do coincide or nearly so for some problems may diverge sharply for others.

The difficulty with the subjective Bayesian approach is that prior knowledge is represented by a probability distribution, and this is more of a commitment than warranted under conditions of partial ignorance. (Uniform or improper priors are just as bad in some respects as anything other sort of prior.) The methods in the (Inference, Inverse) cell all attempt to escape this difficulty by presenting alternative representations of partial ignorance.

Edwards, in particular, uses logarithm of normalized likelihood as a measure of support for a hypothesis. Prior information can be included in the form of a prior support (log likelihood) function; a flat support represents complete prior ignorance.

One place where likelihood methods would deviate sharply from "standard" practice is in a comparison between a sharp and a diffuse hypothesis. Consider H0: X ~ N(0, 100) [diffuse] and H1: X ~ N(1, 1) [standard deviation 10 times smaller]. In standard methods, observing X = 2 would be undiagnostic, since it is not in a sensible tail rejection interval (or region) for either hypothesis. But while X = 2 is not inconsistent with H0, it is much better explained by H1--the likelihood ratio is about 6.2 in favor of H1. In Edwards' methods, H1 would have higher support than H0, by the amount log(6.2) = 1.8. (If these were the only two hypotheses, the Neyman-Pearson lemma would also lead one to a test based on likelihood ratio, but Edwards' methods are more broadly applicable.)

I do not want to appear to advocate likelihood methods. I could give a long discussion of their limitations and of alternatives that share some of their advantages but avoid their limitations. But it is definitely a mistake to dismiss such methods lightly. They are practical (currently widely used in genetics) and are based on a careful and profound analysis of inference.

What is a Meta-Analysis?

Meta-Analysis deals with the art of combining information from the data from different independent sources which are targeted at a common goal. There are plenty of applications of Meta-Analysis in various disciplines such as Astronomy, Agriculture, Biological and Social Sciences, and Environmental Science. This particular topic of statistics has evolved considerably over the last twenty years with applied as well as theoretical developments.

A Meta-analysis deals with a set of RESULTs to give an overall RESULT that is (presumably) comprehensive and valid.

a) Especially when Effect-sizes are rather small, the hope is that one can gain good power by essentially pretending to have the larger N as a valid, combined sample.

b) When effect sizes are rather large, then the extra POWER is not needed for main effects of design: Instead, it theoretically could be possible to look at contrasts between the slight variations in the studies themselves.

If you really trust that "all things being equal" will hold up. The typical "meta" study does not do the tests for homogeneity that should be required

In other words:

1. there is a body of research/data literature that you would like to summarize

2. one gathers together all the admissible examples of this literature (note: some might be discarded for various reasons)

3. certain details of each investigation are deciphered... most important would be the effect that has or has not been found, i.e., how much larger in sd units is the treatment group's performance compared to one or more controls.

4. call the values in each of the investigations in #3.. mini effect sizes.

5. across all admissible data sets, you attempt to summarize the overall effect size by forming a set of individual effects... and using an overall sd as the divisor.. thus yielding essentially an average effect size.

6. in the meta analysis literature... sometimes these effect sizes are further labeled as small, medium, or large....

You can look at effect sizes in many different ways.. across different factors and variables. but, in a nutshell, this is what is done.

I recall a case in physics, in which, after a phenomenon had been observed in air, emulsion data was examined. The theory would have about a 9% effect in emulsion, and behold, the published data gave 15%. As it happens, there was no significant (practical, not statistical) in the theory, and also no error in the data. It was just that the results of experiments in which nothing statistically significant was found were not reported.

This non-reporting of such experiments, and often of the specific results which were not statistically significant, which introduces major biases. This is also combined with the totally erroneous attitude of researchers that statistically significant results are the important ones, and than if there is no significance, the effect was not important. We really need to between the term "statistically significant", and the usual word significant.

It is very important to distinction between statistically significant and generally significant, see Discover Magazine (July, 1987), The Case of Falling Nightwatchmen, by Sapolsky. In this article, Sapolsky uses the example to point out the very important distinction between statistically significant and generally significant: A diminution of velocity at impact may be statistically significant, but not of importance to the falling nightwatchman.

Be careful about the word "significant". It has a technical meaning, not a commonsense one. It is NOT automatically synonymous with "important". A person or group can be statistically significantly taller than the average for the population, but still not be a candidate for your basketball team. Whether the difference is substantively (not merely statistically) significant is dependent on the problem which is being studied.

Meta-analysis is a controversial type of literature review in which the results of individual randomized controlled studies are pooled together to try to get an estimate of the effect of the intervention being studied. It increases statistical power and is used to resolve the problem of reports which disagree with each other. It's not easy to do well and there are many inherent problems.

There is also graphical technique to assess robustness of meta-analysis results. We should carry out the meta-analysis dropping consecutively one study, that is if we have N studies we should do N meta-analysis using N-1 studies in each one. After that we plot these N estimates on the y axis and compare them with a

straight line that represent the overall estimate using all the studies.

Topics in Meta-analysis includes: Odds ratios; Relative risk; Risk difference; Effect size; Incidence rate difference and ratio; Plots and exact confidence intervals.

Further Readings:
Glass, *et al*., *Meta-Analysis in Social Research*, McGraw Hill, 1987
Cooper H., and L. Hedges, (Eds.), *Handbook of Research Synthesis*, Russell Sage Foundation, New York, 1994

Industrial Data Modeling

Industrial Data Modeling is the application of statistical, mathematical and computing techniques to industrial problems. Its applications aimed at science and engineering practitioners and managers in industry, considers the modeling, analysis and interpretation of data in industries associated with science, engineering and biomedicine. The techniques are closely related to those of chemometrics, technometrics and biometrics.

Further Readings:
Montgomery D., and G. Runger, *Applied Statistics and Probability for Engineers*, Wiley, 1998.
Ross Sh., *Introduction to Probability and Statistics for Engineers and Scientists*, Academic Press, 1999.

Prediction Interval

The idea is that if $\bar{x}$ is the mean of a random sample of size n from a normal population, and Y is a single additional observation, then the test statistic $\bar{x}$ - Y is normal with mean 0 and variance (1+ 1/n)s2.

Since we don't actually know s2, we need to use t in evaluating the test statistic. The appropriate Prediction Interval for Y is

$\bar{x}$ ± ta/2.S.(1+1/n)1/2.

This is similar to construction of interval for individual prediction in regression analysis.

Fitting Data to a Broken Line

Fitting data to a broken, how to determine the parameters, a, b, c, and d such that

y = a+ b x, for x less than or equal c
y = a - d c+ (d+ b) x, for x greater than or equal to c

A simple solution is a brute force search across the values of c. Once c is known, estimating a, b, and d is trivial through the use of indicator variables. One may use (x-c) as your independent variable, rather than x, for computational convenience.

Now, just fix c at a fine grid of x values in the range of your data, estimate a, b, and d, and then note what the mean squared error is. Select the value of c that minimizes the mean squared error.

Unfortunately, you won't be able to get confidence intervals involving c, and the confidence intervals for the remaining parameters will be conditional on the value of c.

Further Readings:
For more details, see *Applied Regression Analysis*, by Draper and Smith, Wiley 1981, Chapter 5, section 5.4 on use of dummy variables. example 6.

How to Determine if Two Regression Lines Are Parallel?

Would like to determine if two regression lines are parallel? Construct the following multiple linear regression model:

```
E(y) = b0+ b1X1+ b2X2+ b3X3

 where  X1 = interval predictor variable, X2 = 1 if group 1,                    0 if group 0,  and X3 = X1.X2  Then, E(y|group=0) = b0+
```

That is, E(y|group=1) is a simple regression with a potentially different slope and intercept compared to group=0.

Ho: slope(group 1) = slope(group 0) is equivalent to Ho: b3=0

Use t-test from variables-in-the equation table to test this hypothesis.

Constrained Regression Model

If you fit a regression forcing the intercept to be zero, the standard error of the slope is less. That seems counter-intuitive. The intercept should be included in the model because it is significant, so why is the standard error for the slope in the worse-fitting model actually smaller?

I agree that it's initially counter-intuitive (see below), but here are two reasons why it's true. The variance of the slope estimate for the constrained model is s2 / SXi 2), where Xi are actual X values ands2 is estimated from the residuals. The variance of the slope estimate for the unconstrained model (with intercept) is s2 / Sxi 2), where xi are deviations from the mean, and s2 is still estimated from the residuals). So, the constrained model can have a larger s2 (mean square error/"residual" and standard error of estimate) but a smaller standard error of the slope because the denominator is larger.

r2 also behaves very strangely in the constrained model; by the conventional formula, it can be negative; by the formula used by most computer packages, it is generally larger than the unconstrained r2because it is dealing with deviations from 0, not deviations from the mean. This is because, in effect, constraining

the intercept to 0 forces us to act as if the mean of X and the mean of Y both were 0.

Once you recognize that the s.e. of the slope isn't really a measure of overall fit, the result starts to make a lot of sense. Assume that all your X and Y are positive. If you're forced to fit the regression line through the origin (or any other point) there will be less "wiggle" in how you can fit the line to the data than there would be if both "ends" could move.

Consider a bunch of points that are ALL way out, far from zero, then if you Force the regression through zero, that line will be very close to all the points, and pass through origin, with LITTLE ERROR. And little precision, and little validity. Therefore, no-intercept model is hardly ever appropriate.

Semiparametric and Non-parametric modeling

Many parametric regression models in applied science have a form like response = function(X1,..., Xp, unknown influences). The "response" may be a decision (to buy a certain product), which depends on p measurable variables and an unknown reminder term. In statistics, the model is usually written as

Y = m(X1,..., Xp)+ e

and the unknown e is interpreted as error term.

The most simple model for this problem is the linear regression model, an often used generalization is the Generalized Linear Model (GLM)

Y= G(X1b1+...+ Xpbp)+ e

where G is called the link function. All these models lead to the problem of estimating a multivariate regression. Parametric regression estimation has the disadvantage, that by the parametric "form" certain properties of the resulting estimate are already implied.

Nonparametric techniques allow diagnostics of the data without this restriction. However, this requires large sample sizes and causes problems in graphical visualization. Semiparametric methods are a compromise between both: they support a nonparametric modeling of certain features and profit from the simplicity of parametric methods.

Further Readings:
Härdle W., S. Klinke, and B. Turlach, *XploRe: An Interactive Statistical Computing Environment*, Springer, New York, 1995.

Moderation and Mediation

"Moderation" is an interactional concept. That is, a moderator variable "modifies" the relationships between two other variables. While "Mediation" is a "causal modeling" concept. The "effect" of one variable on another is "mediated" through another variable. That is, there is no "direct effect", but rather an "indirect effect."

Discriminant and Classification

Classification or discrimination involves learning a rule whereby a new observation can be classified into a pre-defined class. Current approaches can be grouped into three historical strands: statistical, machine learning and neural network. The classical statistical methods make distributional assumptions. There are many others which are distribution free, and which require some regularization so that the rule performs well on unseen data. Recent interest has focused on the ability of classification methods to be generalized.

We often need to classify individuals into two or more populations based on a set of observed "discriminating" variables. Methods of classification are used when discriminating variables are:

1. quantitative and approximately normally distributed;
2. quantitative but possibly nonnormal;
3. categorical; or
4. a combination of quantitative and categorical.

It is important to know when and how to apply linear and quadratic discriminant analysis, nearest neighbor discriminant analysis, logistic regression, categorical modeling, classification and regression trees, and cluster analysis to solve the classification problem. SAS has all the routines you need to for proper use of these classifications. Relevant topics are: Matrix operations, Fisher's Discriminant Analysis, Nearest Neighbor Discriminant Analysis, Logistic Regression and Categorical Modeling for classification, and Cluster Analysis.

For example, two related methods which are distribution free are the k-nearest neighbor classifier and the kernel density estimation approach. In both methods, there are several problems of importance: the choice of smoothing parameter(s) or k, and choice of appropriate metrics or selection of variables. These problems can be addressed by cross-validation methods, but this is computationally slow. An analysis of the relationship with a neural net approach (LVQ) should yield faster methods.

Further Readings:
Cherkassky V, and F. Mulier, *Learning from Data: Concepts, Theory, and Methods*, John Wiley & Sons, 1998.
Denison, D., C. Holmes, B. Mallick, and A.Smith, *Bayesian Methods for Nonlinear Classification and Regression*, Wiley, 2002.

Index of Similarity in Classification

In many natural sciences, such as ecologists one is interested in a notion of similarity. The index of similarity is devised for comparing, e.g., the species diversity between two different samples or communities. Let a be the total number of species in sample1, b is the number of species in sample2, and j is the number of species common to both samples, then the widely used similarity index is the Mountford Index defined as:

$$I = 2J/[2ab - j(a+ b)]$$

A rather computationally involved for determining a similarity index (I) is due to Fisher, where I is the solution to the following equation:

$$eaI+ ebI = 1+ e(a+b-j)I$$

The index of similarity could be used as a "distance" so that the minimum distance corresponds to the maximum similarity.

Further Readings:
Hayek L., and M. Buzas, *Surveying Natural Populations*, Columbia University Press, NY, 1996.

Generalized Linear and Logistic Models

The generalized linear model (GLM) is possibly the most important development in practical statistical methodology in the last twenty years. Generalized linear models provide a versatile modeling framework in which a function of the mean response is "linked" to the covariates through a linear predictor and in which variability is described by a distribution in the exponential dispersion family. These models include logistic regression and log-linear models for binomial and Poisson counts together with normal, gamma and inverse Gaussian models for continuous responses. Standard techniques for analyzing censored survival data, such as the Cox regression, can also be handled within the GLM framework. Relevant topics are: Normal theory linear models, Inference and diagnostics for GLMs, Binomial regression, Poisson regression, Methods for handling overdispersion, Generalized estimating equations (GEEs).

Hre is how to obtain degree of freedom number for the 2 log-likelihood, in a logistic regression. Degrees of freedom pertain to the dimension of the vector of parameters for a given model. Suppose we know that a model ln(p/(1-p))=Bo+ B1x+ B2y+ B3w fits a set of data. In this case the vector B=(Bo,B1, B2, B3) is an element of 4 dimensional Euclidean space, or R4.

Suppose we want to test the hypothesis: Ho: B3=0. We are imposing a restriction on our parameter space. The vector of parameters must be of the form: B'=B=(Bo,B1, B2, 0). This vector is an element of a subspace of R4. Namely, B4=0 or the X-axis. The likelihood ration statistic has the form:

2 log-likelihood = 2 log(maximum unrestricted likelihood/maximum restricted likelihood) =
2 log(maximum unrestricted likelihood)-2 log (maximum restricted likelihood)

Which is unrestricted B vector 4-dimensions or degrees of freedom - restricted B vector 3 dimensions or degrees of freedom = 1 degree of freedom which is the difference vector: B''=B-B'=(0,0,0,B4) [one dimensional subspace of R4.

The standard textbook is *Generalized Linear Models* by McCullagh and Nelder (Chapman & Hall, 1989).

```
   LOGISTIC REGRESSION VAR=x  /METHOD=ENTER y x1 x2 f1ros f1ach f1grade bylocus byses  /CONTRAST (y)=Indicator   /contrast (x1)=indicator
```

Other SPSS Commands:

```
Loglinear                       LOGLINEAR,HILOGLINEAR Logistic Regression              LOGLINEAR,PROBIT
```

SAS Commands:

```
Loglinear                       CATMOD  Logistic Regression            LOGISTIC, CATMOD,PROBIT
```

Further Readings:
Harrell F, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer Verlag, 2001.
Hosmer D. Jr., and S. Lemeshow, *Applied Logistic Regression*, Wiley, 2000.
Katz M., *Multivariable Analysis: A Practical Guide for Clinicians*, Cambridge University Press, 1999.
Kleinbaum D., *Logistic Regression: A Self-Learning Text*, Springer Verlag, 1994.
Pampel F., *Logistic Regression: A Primer*, Sage, 2000.

Survival Analysis

Survival analysis is suited to the examination of data where the outcome of interest is 'time until a specific event occurs', and where not all individuals have been followed up until the event occurs. Survival data arise in a literal form from trials concerning life-threatening conditions, but the methodology can also be applied to other waiting times such as the duration of pain relief.

The methods of survival analysis are applicable not only in studies of patient survival, but also studies examining adverse events in clinical trials, time to discontinuation of treatment, duration in community care before re-hospitalisation, contraceptive and fertility studies etc.

If you've ever used regression analysis on longitudinal event data, you've probably come up against two intractable problems:

Censoring: Nearly every sample contains some cases that do not experience an event. If the dependent variable is the time of the event, what do you do with these "censored" cases?

Time-dependent covariate: Many explanatory variables (like income or blood pressure)change in value over time. How do you put such variables in a regression analysis?

Makeshift solutions to these questions can lead to severe biases. Survival methods are explicitly designed to deal with censoring and time-dependent covariates in a statistically correct way. Originally developed by biostatisticians, these methods have become popular in sociology, demography, psychology, economics, political science, and marketing.

In Short, survival Analysis is a group of statistical methods for analysis and interpretation of survival data. Even though survival analysis can be used in a wide variety of applications (e.g. insurance, engineering, and sociology), the main application is for analyzing clinical trials data. Survival and hazard functions, the methods of estimating parameters and testing hypotheses that are the main part of analyses of survival data. Main topics relevant to survival data analysis are: Survival and hazard functions, Types of censoring, Estimation of survival and hazard functions: the Kaplan-Meier and life table estimators, Simple life tables, Peto's Logrank with trend test and hazard ratios and Wilcoxon test, (can be stratified), Wei-Lachin, Comparison of survival functions: The logrank and Mantel-Haenszel tests, The proportional hazards model: time independent and time dependent covariates, The logistic regression model, and Methods for determining sample sizes.

In the last few years the survival analysis software available in several of the standard statistical packages has experienced a major increment in functionality, and is no longer limited to the triad of Kaplan-Meier curves, logrank tests, and simple Cox models.

Further Readings:
Hosmer D., and S. Lemeshow, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley, 1999.
Janssen P., J. Swanepoel, and N. Veraverbeke, The modified bootstrap error process for Kaplan-Meier quantiles, *Statistics & Probability Letters*, 58, 31-39,

2002.
Kleinbaum D., *et al*., *Survival Analysis: A Self-Learning Text*, Springer-Verlag, New York, 1996.
Lee E., *Statistical Methods for Survival Data Analysis*, Wiley, 1992.
Therneau T., and P. Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer 2000. This book provides thorough discussion on Cox PH model. Since the first author is also the author of the survival package in S-PLUS/R, the book can be used closely with the packages in addition to SAS.

Association Among Nominal Variables

There are many measures of association between two dichotomous variables, such as the odds ratio (AD/BC), Yule's Q = (AD-BC/AD+BC) which is a simple mapping of the odds ratio onto [-1,1], the proportional difference (requires treating one of the variables as "independent" and the other "dependent"), Cramer's V, the contingency coefficient C, the uncertainly coefficient, and the relative risk. Some of those measures may be more appropriate than others for a given situation, however, those based on the odds ratio are easier to interpret. Odds ratios can be thought of as the effect of one outcome on another. If condition 1 is true, what effect has it on the odds of condition 2 being true? Almost all of these statistics are described in the Numerical Recipes, by Press et al.

Spearman's Correlation, and Kendall's tau Application

How would you compare the values of two variables to determine whether they are ordered the same? For example:

```
Var1  Var2 Obs 1   x    x Obs 2   y    z Obs 3   z    y
```

Is Var1 ordered the same as Var2? Two measures are Spearman's rank order correlation, and Kendall's tau.

Further Readings:
For more details see, e.g., *Fundamental Statistics for the Behavioral Sciences*, by David C. Howell, Duxbury Pr., 1995.

Repeated Measures and Longitudinal Data

Repeated measures and longitudinal data require special attention because they involve correlated data that commonly arise when the primary sampling units are measured repeatedly over time or under different conditions. Normal theory models for split-plot experiments and repeated measures ANOVA can be used to introduce the concept of correlated data. PROC GLM and PROC MIXED in the SAS system may be used. Mixed linear models provide a general framework for modeling covariance structures, a critical first step that influences parameter estimation and tests of hypotheses. The primary objectives are to investigate trends over time and how they relate to treatment groups or other covariates. Techniques applicable to non-normal data, such as McNemar's test for binary data, weighted least squares for categorical data, and generalized estimating equations (GEE) are the main topics. The GEE method can be used to accommodate correlation when the means at each time point are modelled using a generalized linear model. Relevant topics are: Balanced split-plot and repeated measures designs, Modeling covariance structures of repeated measures, Repeated measures with unequally spaced times and missing data, Weighted least squares approach to repeated categorical data, Generalized estimating equation (Gee) method for marginal models, Subject-specific versus population averaged interpretation of regression coefficients, and Computer implementation using S-plus and the SAS system. The following describes the McNemar's test for binary data.

McNemar Change Test: For the yes/no questions under the two conditions, set up a 2x2 contingency table:

```
        f11     f10     f01     f00
```

McNemar's test of correlated proportions is z = (f01 - f10)/(f01+ f10)½.

For those items yielding a score on a scale, the conventional t-test for correlated samples would be appropriate, or the Wilcoxon signed-ranks test.

What Is a Systematic Review?

Health care decision makers need to access research evidence to make informed decisions on diagnosis, treatment and health care management for both individual patients and populations. Systematic reviews are recognized as one of the most useful and reliable tools to assist this practice of evidence-based health care. These courses aim to train health care professionals and researchers in the science and methods of systematic reviews.

There are few important questions in health care which can be informed by consulting the result of a single empirical study. Systematic reviews attempt to provide answers to such problems by identifying and appraising all available studies within the relevant focus and synthesizing their results, all according to explicit methodologies. The review process places special emphasis on assessing and maximizing the value of data, both in issues of reducing bias and minimizing random error. The systematic review method is most suitably applied to questions of patient treatment and management, although it has also been applied to answer questions regarding the value of diagnostic test results, likely prognoses and the cost-effectiveness of health care.

Information Theory

Information theory is a branch probability and mathematical statistics that deal with communication systems, data transmission, cryptography, signal to noise ratios, data compression, etc. Claude Shannon is the father of information theory. His theory considered the transmission of information as a statistical phenomenon and gave communications engineers a way to determine the capacity of a communication channel about the common currency of bits

Shannon defined a measure of entropy as:

$$H = - S p i \log pi,$$

that, when applied to an information source, could determine the capacity of the channel required to transmit the source as encoded binary digits. Shannon's measure of entropy is taken as a measure of the information contained in a message. This is unlike to the portion of the message that is strictly determined (hence predictable) by inherent structures.

Entropy as defined by Shannon is closely related to entropy as defined by physicists in statistical thermodynamics. This work was the inspiration for adopting the term entropy in information theory. Other useful measures of information include mutual information which is a measure of the correlation between two event sets. Mutual information is defined for two events X and Y as:

$$M(X, Y) = H(X, Y) - H(X) - H(Y)$$

where H(X, Y) is the join entropy defined as:

$$H(X, Y) = - S\, p(xi, yi) \log p(xi, yi),$$

Mutual information is closely related to the log-likelihood ratio test for multinomial distribution, and to Pearson's Chi-square test.

The field of Information Science has since expanded to cover the full range of techniques and abstract descriptions for the storage, retrieval and transmittal of information.

Incidence and Prevalence Rates

Incidence rate (IR) is the rate at which new events occur in a population. It is defined as: Number of new events in a specified period divided by Number of persons exposed to risk during this period

Prevalence rate (PR) measures the number of cases that are present at a specified period of time. It is defined as: Number of cases present at a specified period of time divides by Number of persons at risk at that specified time.

These two measures are related when considering the average duration (D). That is, PR = IR. D

Note that, for example, county-specific disease incidence rates can be unstable due to small populations or low rates. In epidemiology one can say that IR reflects probability to Become thick at given age, while the PR reflects probability to Be thick at given age.

Other topics in clinical epidemiology include the use of receiver operator curves, and the sensitivity, specificity, predictive value of a test.

Further Readings:

Kleinbaum D., L. Kupper, and K. Muller, *Applied Regression Analysis and Other Multivariable Methods*, Wadsworth Publishing Company, 1988.
Kleinbaum D., *et al*., *Survival Analysis: A Self-Learning Text*, Springer-Verlag, New York, 1996.
Miettinen O., *Theoretical Epidemiology*, Delmar Publishers, 1986.

Software Selection

The availability of personal computer, computational software, and visual representations of data enables the managers to concentrate on the revealing useful facts from figures. Since the burden of computation has been eliminated, the managers are now able to focus on probing issues and search for creative decision-making under uncertainty. However, you have to be careful when selecting a statistical software. A short list of item for comparison is:

1) Ease of learning,
2) Amount of help incorporated for the user,
3) Level of the user,
4) Number of tests and routines involved,
5) Ease of data entry,
6) Data validation (and if necessary, data locking and security),
7) Accuracy of the tests and routines,
8) Integrated data analysis (graphs and progressive reporting on analysis in one screen),
9) Cost

No one software meets everyone's needs. Determine the needs first and then ask the questions relevant to the above seven criteria.

Spatial Data Analysis

Data that is geographically or spatially referenced is encountered in a very wide variety of practical contexts. In the same way that data collected at different points in time may require specialised analytical techniques, there are a range of statistical methods devoted to the modelling and analysis of data collected at different points in space. Increased public sector and commercial recording and use of data which is geographically referenced, recent advances in computer hardware and software capable of manipulating and displaying spatial relationships in the form of digital maps, and an awareness of the potential importance of spatial relationships in many areas of research, have all combined to produced an increased interest in spatial analysis. Spatial Data Analysis is concerned with the study of such techniques---the kind of problems they are designed to address, their theoretical justification, when and how to use them in practice.

Many natural phenomena involve a random distribution of points in space. Biologists who observe the locations of cells of a certain type in an organ, astronomers who plot the positions of the stars, botanists who record the positions of plants of a certain species and geologists detecting the distribution of a rare mineral in rock are all observing spatial point patterns in two or three dimensions. Such phenomena can be modelled by spatial point processes.

The spatial linear model is fundamental to a number of techniques used in image processing, for example, for locating gold/ore deposits, or creating maps. There are many unresolved problems in this area such as the behavior of maximum likelihood estimators and predictors, and diagnostic tools. There are strong connections between kriging predictors for the spatial linear model and spline methods of interpolation and smoothing. The two-dimensional version of splines/kriging can be used to construct deformations of the plane, which are of key importance in shape analysis.

For analysis of spatially auto-correlated data in of logistic regression for example, one may use of the Moran Coefficient which is available is some statistical packages such as Spacestat. This statistic tends to be between -1 and+1, though are not restricted to this range. Values near+1 indicate similar values tend to cluster; values near -1 indicate dissimilar values tend to cluster; values near -1/(n-1) indicate values tend to be randomly scattered.

Boundary Line Analysis

The boundary line analysis is dealing with developing the analytical syntheses of real property law, land surveying procedures, & scenario development which helps with decisions for the development of most probable scenarios of boundary location.

The main application of this analysis is in the soil electrical conductivity (EC) which stems from the fact that sands have a low conductivity, silts have a medium conductivity and clays have a high conductivity. Consequently, conductivity (measured at low frequencies) correlates strongly to soil grain size and texture.

The boundary line analysis, therefore, is a method of analyzing yield with soil electrical conductivity data. This method isolates the top yielding points for each soil EC range and fits a non-linear line or equation to represent the top-performing yields within each soil EC range. This method knifes through the cloud of EC/Yield data and describes their relationship when other factors are removed or reduced. The upper boundary represents the maximum possible response to that limiting factor, (e.g. EC), and points below the boundary line represents conditions where other factors have limited the response variable. Therefore, one may also use boundary line analysis to compare responses among species.

Further Readings:
Kitchen N., K Sudduth, and S. Drummond, Soil Electrical Conductivity as a Crop Productivity Measure for Claypan Soils, *Journal of Production Agriculture*, 12(4), 607-617, 1999.

Geostatistics Modeling

The Geostatistics modeling combines the classical statistics-based techniques with space/time imaging. The modeling process includes a group of spatiotemporal concepts and methods that are based on stochastic data analysis. The aim of such modeling approach is to provide a deeper understanding of a theory of knowledge prior to development of mathematical models of scientific mapping and imaging across space and time. One effective approach is the to provides a fundamental insight into the mapping problem in which the knowledge of a natural variable, not the variable itself, is the direct object of study. Several well-known models in this category include the spatiotemporal random fields such as space/time fractals and wavelets which are special cases of the generalized random field modeling.

Further Readings:
Christakos G., *Modern Spatiotemporal Geostatistics*, Oxford University Press, 2000.

Box-Cox Power Transformation

In certain cases data distribution is not normal (Gaussian), and we wish to find the best transformation of variable in order to obtain a Gaussian data distribution for further statistical processing.

Among others the Box-Cox power transformation is often used for this purpose.

```
y = (xp - 1)/p, for p not zero  y = log x,      for p = 0
```

trying different values of p between -3 and+3 is usually sufficient but there are MLE methods for estimating the best p. A good source on this and other transformation methods is
Madansky A., *Prescriptions for working Statisticians*, Springer-Verlag, 1988.

For percentages or proportions (such as for binomial proportions), Arcsine transformations would work better. The original idea of Arcsin(p½)is to establish variances as equal for all groups. The arcsin transform is derived analytically to be the variance-stabilizing and normalizing transformation. The same limit theorem also leads to the square root transform for Poisson variables (such as counts) and to the arc hyperbolic tangent (i.e., Fisher's Z) transform for correlation. The Arcsin Test yields a z and the 2x2 contingency test yields a chi-sq. But z2 = chi-sq, for large sample size. A good source is
Rao C., *Linear Statistical Inference and Its Applications*, Wiley, 1973.

How to normalize a set of data consisting of negative and positive values, and make them positive between the range 0.0 to 1.0? Define XNew = (X-min)/(max-min).

Box & Cox power transformation is also very effective for a wide variety of nonnormality:

y(transformed) = yl

where l ranges (in practice) from -3.0 to+3.0. As such it includes, inverse, square root, logarithm, etc. Note that as l approaches 0, one gets a log transformation.

Multiple Comparison Tests

Duncan's multiple-range test: This is one of the many multiple comparison procedures. It is based on the standardized range statistic by comparing all pairs of means while controlling the overall Type I error at a desirable level. While it does not provide interval estimates of the difference between each pair of means, however, it does indicate which means are significantly different from the others. For determining the significant differences between *a single control group* mean and the other means, one may use the Dunnett's multiple-comparison test.

Multiple comparison procedures include topics such as Control of the family-Wise Error rate, The closure Principle, Hierarchical Families of Hypotheses, Single-Step and Stepwise Procedures, and P-value Adjustments. Areas of applications include multiple comparisons among treatment means, multiple endpoints in clinical trials, multiple sub-group comparisons, etc.

Nemenyi's multiple comparison test is analogous to Tukey's test, using rank sums in place of means and using [n2k(nk+1)/12]½ as the estimate of standard error (SE), where n is the size of each sample and k is the number of samples (means). Similarly to the Tukey test, you compare (rank sum A - rank sum B)/SE to the studentized range for k. It is also equivalent to the Dunn/Miller test which uses mean ranks and standard error [k(nk+1)/12]½.

Multilevel Statistical Modeling: The two widely used software packages are MLwiN and winBUGS. They perform multilevel modeling analysis and analysis of hierarchical datasets, Markov chain Monte Carlo (MCMC) methodology and Bayesian approaches.

Further Readings:
Liao T., *Statistical Group Comparison*, Wiley, 2002.

Antedependent Modeling for Repeated Measurements

Repeated measures data arise when observations are taken on each experimental unit on a number of occasions, and time is a factor of interest.

Many techniques can be used to analyze such data. Antedependence modeling is a recently developed method which models the correlation between observations at different times.

Split-half Analysis

What is split-half analysis? Split your sample in half. Factor analyses each half. Do they come out the same (or similar) as each other? Alternatively (or also), take more than two 2 random subsample of your sample and do the same.

Notice that this is (like factor analysis itself) an "exploratory", not inferential technique, i.e. hypothesis testing, confidence intervals etc. simply do not apply.

Alternatively, randomly split the sample in half and then do an exploratory factor analysis on Sample 1. Use those results to do a confirmatory factor analysis with Sample 2.

Sequential Acceptance Sampling

Acceptance sampling is a quality control procedure used when a decision on the acceptability of the batch has to be made from tests done on a sample of items from the batch.

Sequential acceptance sampling minimizes the number of items tested when the early results show that the batch clearly meets, or fails to meet, the required standards.

The procedure has the advantage of requiring fewer observations, on average, than fixed sample size tests for a similar degree of accuracy.

Local Influence

Cook's distance measures the effect of removing a single observation on regression estimates. This can be viewed as giving an observation a weight of either zero or one: local influence allows this weight to be small but non-zero.

Cook defined local influence in 1986, and made some suggestions on how to use or interpret it; various slight variations have been defined since then. But problems associated with its use have been pointed out by a number of workers since the very beginning.

Variogram Analysis

Variables are often measured at different locations. The patterns in these spatial variables may be extrapolated by variogram analysis.

A variogram summarizes the relationship between the variance of the difference in pairs of measurements and the distance of the corresponding points from each other.

Credit Scoring: Consumer Credit Assessment

Credit Scoring is now in widespread use across the retail credit industry. At its simplest, a credit scorecard is a model usually statistical, but in use it is embedded in a computer and or human process.

Accurate assessment of financial exposure is vital for continued business success. Accurate, and usable information are essential for good credit assessment in commercial decision making. The consumer credit environment is in a state of great change, driven by developments in computer technology, more demanding customers, availability of new products and increased competition. Banks and other financial institutions are coming to rely more and more on increasingly sophisticated mathematical and statistical tools. These tools are used in a wide range of situations, including predicting default risk, estimating likely profitability, fraud detection, market segmentation, and portfolio analysis. The credit card market as an example, has changed the retail banking industry, and consumer loans.

Both the tools, the behavioral scoring, and the characteristics of consumer credit data are usually the bases for a good decision. The statistical tools include linear and logistic regression, mathematical programming, trees, nearest neighbor methods, stochastic process models, statistical market segmentation, and neural networks. These techniques are used to assess and predict consumers credit scoring.

Further Readings:
Lewis E., *Introduction to Credit Scoring*, Fair, Isaac & Co., 1994. Provides a general introduction to the issues of building a credit scoring model.

Components of the Interest Rates

The interest rates as quoted in the newspapers and by banks consist of several components. The most important three are:

The pure rate: This is the time value of money. A promise of 100 units next year is not worth 100 units this year.

The price-premium factor: If prices go up 5% each year, interest rates go up at least 5%. For example, under the Carter Administration, prices rose about 15% per year for a couple of years, interest was around 25%. Same thing during the Civil War. In a deflationary period, prices may drop so this term can be negative.

The risk factor: A junk bond may pay a larger rate than a treasury note because of the chance of losing the principal. Banks in a poor financial condition must pay higher rates to attract depositors for the same reason. Threat of confiscation by the government leads to high rates in some countries.

Other factors are generally minor. Of course, the customer sees only the sum of these terms. These components fluctuate at different rates themselves. This makes it hard to compare interest rates across disparate time periods or economic condition. The main questions are: how are these components combined to form the index? A simple sum? A weighted sum? In most cases the index is form both empirically and assigned on basis of some criterion of importance. The

same applies to other index numbers.

Partial Least Squares

Partial Least Squares (PLS) regression is a multivariate data analysis technique which can be used to relate several response (Y) variables to several explanatory (X) variables.

The method aims to identify the underlying factors, or linear combination of the X variables, which best model the Y dependent variables.

Growth Curve Modeling

Growth is a fundamental property of biological systems, occurring at the level of populations, individual animals and plants, and within organisms. Much research has been devoted to modeling growth processes, and there are many ways of doing this: mechanistic models, time series, stochastic differential equations etc.

Sometimes we simply wish to summarize growth observations in terms of a few parameters, perhaps in order to compare individuals or groups. Many growth phenomena in nature show an "S" shaped pattern, with initially slow growth speeding up before slowing down to approach a limit.

These patterns can be modelled using several mathematical functions such as generalized logistic and Gompertz curves.

Saturated Model & Saturated Log Likelihood

A saturated model is usually one that has no residual df. What is a "saturated" log likelihood? So the "saturated LL" is the LL for a saturated model. It is often used when comparisons made between the log likelihood with an intercept only and the log likelihood for a particular model specification.

Pattern recognition and Classification

Pattern recognition and classification are fundamental concepts for understanding living systems and essential for realizing artificial intelligent systems. Applications include 3D modelling, motion analysis, feature extraction, device positioning and calibration, feature recognition, solutions to classification problems to industrial and medical applications.

What is Biostatistics?

Biostatistics is a subdiscipline of Statistics which focuses on statistical support for the areas of medicine, environmental science, public health, and related fields. Practitioners span the range from the very applied to the very theoretical. The information which is useful to the biostatistician spans the range from that needed by a general statistician, to more subject-specific scientific details, to ordinary information that will improve communication between the biostatistician and other scientists and researchers.

Recent advancement in human genome marks a major step in the advancement of understanding how the human body works at a molecular level. The biomedical statistics identifies the need for computational statistical tools to meet important challenges in biomedical studies. The active areas are:

- Clustering of very large dimensional data such as the micro-array.
- Clustering algorithms that support biological meaning.
- Network models and simulations of biological pathways.
- Pathway estimation from data.
- Integration of multi-format and multi-type data from heterogeneous databases.
- Information and knowledge visualization techniques for biological systems.

Further Reading:

Cleophas T., A. Zwinderman, and T. Cleophas, *Statistics Applied to Clinical Trials*, Kluwer Academic Publishers, 2002.
Zhang W., and I. Shmulevich, *Computational and Statistical Approaches to Genomics*, Kluwer Academic Publishers, 2002.

Evidential Statistics

Statistical methods aim to answer a variety of questions about observations. A simple example occurs when a fairly reliable test for a condition C, has given a positive result. Three important types of questions are:

1. Should this observation lead me to believe that condition C is present?
2. Does this observation justify my acting as if condition C were present?
3. Is this observation evidence that condition C is present?

We must distinguish among these three questions in terms of the variables and principles that determine their answers. Questions of the third type, concerning the "evidential interpretation" of statistical data, are central to many applications of statistics in many fields.

It is already recognized that for answering the evidential question current statistical methods are seriously flawed which could be corrected by a applying the the Law of Likelihood. This law suggests how the dominant statistical paradigm can be altered so as to generate appropriate methods for objective, quantitative representation of the evidence embodied in a specific set of observations, as well as measurement and control of the probabilities that a study will produce weak or misleading evidence.

Further Reading:

Royall R., *Statistical Evidence: A Likelihood Paradigm*, Chapman & Hall, 1997.

Statistical Forensic Applications

Cases abound about the role if evidence and inference in constructing and testing arguments and this can be best seen in police and lawyer training where there has been little if any formal instruction on the structural and temporal elements of evidential reasoning. However, little sign exists of methodological approaches to organising evidence and thought as well as a lack of awareness of the benefits such an approach can bring. In addition, there is little regard for the way in which evidence has to be discovered, analyzed and presented as part of a reasoned chain or argument.

One consequence of the failure to recognize the benefits that an organized approach can bring is our failure to move evidence as a discipline into volume case analytics. Any cursory view of the literature reveals that work has centered on thinking about single cases using narrowly defined views of what evidential reasoning involves. There has been an over emphasis on the formal rules of admissibility rather than the rules and principles of a methodological scientific approach.

As the popularity of using DNA evidence increases, both the public and professionals increasingly regard it as the last word on a suspect's guilt or innocence. As citizens go about their daily lives, pieces of their identities are scattered in their wake. It could as some critics warn, one day place an innocent person at the scene of a crime.

The traditional methods of statistical forensic, for example, for facial reconstruction date back to the Victorian Era. Tissue depth data was collected from cadavers at a small number of landmark sites on the face. Samples were tiny, commonly numbering less than ten. Although these data sets have been superceded recently by tissue depths collected from the living using ultrasound, the same twenty-or-so landmarks are used and samples are still small and under-representative of the general population. A number of aspects of identity--such as age, height, geographic ancestry and even sex--can only be estimated from the skull. Current research is directed at the recovery of volume tissue depth data from magnetic resonance imaging scans of the head of living individuals; and the development of simple interpolation simulation models of obesity, ageing and geographic ancestry in facial reconstruction.

Further Reading:
Gastwirth J., (Ed.), *Statistical Science in the Courtroom*, Springer Verlag, 2000.

Spatial Statistics

Many natural phenomena involve a random distribution of points in space. Biologists who observe the locations of cells of a certain type in an organ, astronomers who plot the positions of the stars, botanists who record the positions of plants of a certain species and geologists detecting the distribution of a rare mineral in rock are all observing spatial point patterns in two or three dimensions. Such phenomena can be modelled by spatial point processes.

Further Readings:
Diggle P., *The Statistical Analysis of Spatial Point Patterns*, Academic Press, 1983.
Ripley B., *Spatial Statistics*, Wiley, 1981.

What Is the Black-Sholes Model?

The benchmark theory of statistical model for option pricing derivative and evaluation is the Black-Sholes-Merton theory (the Black-Sholes model is a special case which is the limiting distribution of the binomial model), based on Brownian motion as the driving noise process for stock prices. In this model the distributions of financial returns of the stocks in a portfolio are multivariate normal. There are certain limitations in this model, which are, e.g., symmetry and thin tails, which are not the characteristics of the real data. The One may use the Barndorff-Nielsen generalized hyperbolic family, which includes the normal variance-mean mixtures instead of pure multivariate normal.

Further Readings:
Clewlow L., and C. Strickland, *Implementing Derivatives Models*, John Wiley & Sons, 1998.

What Is a Classification Tree

Basically for each variable, all values are checked and a measure of purity calculated, i.e., loosely the number of classification errors is quantified. The value and variable with lowest split is chosen as the node. This process can then be repeated until all distinct combination of values of independent values have been found. Unfortunately the resulting tree over-fits the data, and would not be vary good for new data sets.

There are several methods of deciding when to stop. The simplest method is to split the data into two samples. A tree is developed with one sample and tested with another. The mis-classification rate is calculated by fitting the tree to the test data set and increasing the number of branches one at a time. As the number of nodes used changes the mis-classification rate changes. The number of nodes which minimize the mis-classification rate is chosen.

Graphical Tools for High-Dimensional Classification: Statistical algorithmic classification methods include techniques such as trees, forests, and neural nets. Such methods tend to share two common traits. They can often have far greater predictive power than the classical model-based methods. And they are frequently so complex as to make interpretation very difficult, often resulting in a "black box" appearance. An alternative approach is using graphical tool to facilitate investigation of the inner workings of such classifiers. The A generalization of the ideas such as the data image, and the color histogram allows simultaneous examination of dozens to hundreds of variables across similar numbers of observations. Additional information can be visually incorporated as to true class, predicted class, and casewise variable importance. Careful choice of orderings across cases and variables can clearly indicate clusters, irrelevant or redundant variables, and other features of the classifier, leading to substantial improvements in classifier interpretability.

The various programs vary in how they operate. For making splits, most programs use definition of purity. More sophisticated methods of finding the stopping rule have been developed and depend on the software package.

What Is a Regression Tree

A regression tree is like a classification tree, only with a continuous target (dependent) variable. Prediction of target value for a particular case is made by assigning that case to a node (based on values for the predictor variables) and then predicting the value of the case as the mean of its node (sometimes adjusted for priors, costs, etc.).

The Tree-based models known also as recursive partitioning have been used in both statistics and machine learning. Most of their applications to date have, however, been in the fields of regression, classification, and density estimation.

S-PLUS statistical package includes some nice features such as non-parametric regression and tree-based models.

Further Readings:
Breiman L., J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, CRC Press, Inc., Boca Raton, Florida, 1984.

Cluster Analysis for Correlated Variables

The purpose of Cluster sampling is typically to:

- characterize a specific group of interest,
- compare two or more specific groups,
- discover a pattern among several variables.

Cluster analysis is used to classify observations with respect to a set of variables. The widely used Ward's method is predisposed to find spherical clusters and may perform badly with very ellipsoidal clusters generated by highly correlated variables (within clusters).

To deal with high correlations, some model-based methods are implemented in the S-Plus package. However, a limitation of their approach is the need to assume the clusters have a multivariate normal distribution, as well as the need to decide in advance what the likely covariance structure of the clusters is.

Another option is to combine the principal component analysis with cluster analysis.

Further Readings:
Baxter M., *Exploratory Multivariate Analysis in Archaeology*, pp. 167-170, Edinburgh University Press, Edinburgh, 1994.
Manly F., *Multivariate Statistical Methods: A Primer*, Chapman and Hall, London, 1986.

Capture-Recapture Methods

Capture-recapture methods were originally developed in the wildlife biology to estimate the population size of some species of wild animals.

Tchebysheff Inequality and Its Improvements

The Tchebysheff's inequality is often used to put bounds on the probability that proportion of random variable X will be within k > 1 standard deviation of the mean mu for any probability distribution. In other words:

$$P [|X - m| \geq k s] \leq 1/k2, \text{ for any } k > 1$$

The symmetric property of Tchebysheff's inequality is useful, e.g., in constructing control limits is the quality control process. However the limits are very conservative because of lack of knowledge about the underlying distribution. This bounds can be improved (i.e., becomes tighter) if we have some knowledge about the population distribution. For example, if the population is homogeneous, that is its distribution is unimodal, then,

$$P [|X - m| \geq k s] \leq 1/(2.25k2), \text{ for any } k > 1$$

The above inequality is known as the Camp-Meidell inequality.

Further Readings:

Efron B., and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall (now the CRC Press), 1994. Contains a test for multimodality that is based on the Gaussian kernel density estimates and then test for multimodality by using the window size approach.
Grant E., and R. Leavenworth, *Statistical Quality Control*, McGraw-Hill, 1996.
Ryan T., *Statistical Methods for Quality Improvement*, John Wiley & Sons, 2000. A very good book for a starter.

Frechet Bounds for Dependent Random Variables

The simplest form of the Frechet bounds for two dependent random variables A and B with known marginal probability P(A), and P(B), respectively is:

$$\max[0, P(A)+ P(B) - 1] \leq P(A \text{ and } B) \leq \min[P(A), P(B)]$$

Frechet Bounds is often used in stochastic processes with the effect of dependencies, such as estimating an upper and/or a lower bound on the queue length in a queuing system with two different but known marginal inter-arrivals times distributions of two types of customers.

Statistical Data Analysis in Criminal Justice

This topic usually refers to the wide range of statistics used in the criminal justice system. For example, statistical analysis of the issue of how much crime is drug-related by using the available criminal justice databases, and other source of data. The main issue for the statisticians is to access the specific unit record files for secondary analysis and the long-term implications for evidence based policy making. These analyses must be carried out usually within the specific criminal justice system considering the existence of limitations such as the ethical norms on data release and legislation on privacy and confidentiality.

Further Readings:
McKean J., and Bryan Byers, *Data Analysis for Criminal Justice and Criminology*, Allyn & Bacon, 2000.
Walker J., *Statistics in Criminal Justice: Analysis and Interpretation*, Aspen Publishers, Inc., 1999.

What is Intelligent Numerical Computation?

There exist a few computer algebra program software in the market that solve several numerical problem types, which can not be solved by using the ordinary numerical methods. The technique mostly used is to transform the problems, which are difficult to be solved through the ordinary methods, to equivalent problems but are easy to be solved, by defining measure functions that assess the suitable method for every type of problems. The aim of such software is to make the students able to use this package, rather than writing their own programs in any other programming languages.

Software Engineering by Project Management

Software Engineering by Project management Techniques aims at the capital risks on projects to be evaluated, and calculates the financial contingency required to cover those risks in a rational and defensible manner to make bug-free software in a systematic approach. Too often the project contingency is guesstimated as a "gut feel" amount, without much consideration for the real risks involved. The technique enables disciplined estimating, and calculates the required contingency using the proven statistical method known such as Monte Carlo experimentation.

The software project scheduling and tracking is to create a network of software engineering tasks that will enable you to get the job done on time. Once the network is created, you have to assign responsibility for each task, make sure it gets done, and adapt the network as risks become reality.

Further Readings:
Ricketts I., *Managing Your Software Project: A Student's Guide*, London, Springer, 1998.

Chi-Square Analysis for Categorical Grouped Data

Suppose you have the summary data for each categories rather than raw data, and you wish to perform the Chi-Square test, that is when one only has the cell data, not the data from each individual. As a numerical example, consider the following data set:

Group	Yes	Uncertain	No
1	10	21	23
2	12	15	18

One may first construct an equivalent alternative categorical table as follows:

Group	Reply	Count
1	Y	10
1	U	21
1	N	23
2	Y	12
2	U	15
2	N	18

Now, weight the data by counts and then perform the Chi-Square analysis.

Further Reading:
Agresti A., *Categorical Data Analysis*, Wiley, 2002.
Kish R., G. Kalton, S. Heeringa, C. O'Muircheartaigh, and J. Lepkowski, *Collected Papers of Leslie Kish*, Wiley, 2002.

Cohen's Kappa: A Measures of Data Consistency

Cohen's kappa measures the agreement internal consistency based on a contingency table. In this context a measure of agreement assesses the extent to which two raters give the same ratings to the same objects. The set of possible values for one rater forms the columns and the same set of possible values for some second rater forms the rows.

Kappa k = [observed concordance - concordance by chance]/[1- concordance by chance]

Where "by chance" is calculated as in chi-square: multiply row marginal times column marginal and divide by n.

One may use this measure as a decision-making tool:

Kappa k	Interpretation
k < 0.00	Poor
0.00 £ k < 0.20	Slight
0.20 £ k < 0.40	Fair
0.40 £ k < 0.60	Moderate
0.60 £ k < 0.80	Substantial
0.80 £ k	Almost Perfect

This interpretation is widely accepted, and many scientific journals routinely publish papers using this interpretation for the result of test of hypothesis.

Further Reading:
Looney S., *Biostatistical Methods*, (ed.), Humana Press, 2002.
Rust R., and B. Cooil, Reliability measures for qualitative data: Theory and implications, *Journal of Marketing Research*, 31(1), 1-14, 1994.

Modeling Dependent Categorical Data

One may apply regression models to the categorical dependent variables. However, due to the non-linearities of these models the statistical analysis and interpretation of these models is not an easy task. still difficult The most premising approach is via the method of maximum likelihood estimation in developing the logit and probit models for binary and ordinal data. The multinomial logit model is often used for nominal data. An extensions of modeling for count data, includes modeling process for Poisson regression, negative binomial regression, and the zero modified models.

Further Readings:
Agresti A., *An Introduction to Categorical Data Analysis*, Wiley, 1996.

The Deming Paradigm

While the common practice of Quality Assurance aims to prevent bad units from being shipped beyond some allowable proportion, Statistical Process Control (SPC) ensures that bad units are not created in the first place. Its philosophy of continuous quality improvement, to a great extent responsible for the success of Japanese manufacturing, is rooted in a paradigm as process-oriented as physics, yet produces a friendly and fulfilling work environment.

Further Reading:
Thompson J., and J. Koronacki, *Statistical Process Control: The Deming Paradigm and Beyond*, CRC Press, 2001.

Reliability & Repairable System

Reliability modeling uses subjective judgements to construct models at many different levels. One area is in the construction of joint probability distributions for the lifetime of several pieces of equipment, or for the failure times due to different failure modes of a single piece of equipment. When there is good reason to believe given marginal distributions for the failure times, the problem of selecting a marginal distribution is equivalent to that of selecting a copula. In other situations identification of the copula alone is important, for example in competing risk where the copula together with competing risk data enable identification of the full joint distribution.

The primary intent of reliability engineering is to improve reliability, and almost all systems of interest to reliability engineers are designed to be repairable, this is the most important reliability concept. It is also the simplest, in sharp contrast, spacing between order statistics of the times to failure of non-repairable items (i.e., parts) eventually become stochastically larger. Even under any physically plausible model of wearout. Moreover, if parts are put on test simultaneously and operated continuously, the spacing between order statistics, which are times between failures, occur exactly in calendar time. Because of non-zero repair times, this is never exactly true for a repairable system. As long as a system is non-repairable, the focus usually should be on the underlying distribution's hazard function. Correspondingly, if it is repairable, the focus usually should be on the underlying process's intensity function. However, even though hazard and intensity functions can be - and sometimes have to be - represented by the same mathematical function, the differences in interpretation are significantly different.

Further Reading:
Ascher H., and H. Feingold, *Repairable Systems Reliability: Modeling, Inference, Misconceptions and Their Causes*, Marcel Dekker, 1984.

Computation of Standard Scores

In many areas such as education and psychology, it is often desired to convert test scores (called *raw scores*) to *standard scores* (scores in standard units) with a predetermined mean and standard deviation. This may be accomplished as follows:

$$z = \frac{\sigma'}{\sigma} \times (X - \mu) + \mu'$$

where m = raw score mean

s = raw score standard deviation

X = raw score

m ¢ = new mean

s ¢ = new standard deviation

Suppose a population of psychological test scores has a mean of 70 and a standard deviation of 8 and it is desired to convert these scores to standard scores with a mean of 100 and a standard deviation of 20. If 40 is one of the raw scores in the population, we may apply the foregoing equation to convert this to a standard score by substituting

m = 70, s = 8, X = 40, m ¢ = 100, s ¢ = 20 to obtain

$$Z = \frac{20}{08} \times (40 - 70) + 100 = 25$$

Quality Function Deployment (QFD)

A number of activities must be conducted when carrying out QFD. Some of the typical activities are listed as follows:

1. Analyzing customer requirements.
2. Identifying design features.
3. Establishing interactions between customer requirements and design features.
4. Carrying out competitive benchmarking in technical and/or market terms.
5. Analyzing the results and deriving implications.

A roadmap with the format and procedure is often used to guide the analyst through these steps and record the results obtained. This roadmap is called the QFD worksheet.

Further Readings:
Franceschini F., *Advanced Quality Function Deployment*, St. Lucie Press, 2002.

Event History Analysis

Sometimes data on the exact time of a particular event (or events) are available, for example on a group of patients. Examples of events could include asthma attach; epilepsy attack; myocardial infections; hospital admissions. Often, occurrence (and non-occurrence) of an event is available on a regular basis (e.g., daily) and the data can then be thought of as having a repeated measurements structure. An objective may be to determine whether any concurrent events or

measurements have influenced the occurrence of the event of interest. For example, daily pollen counts may influence the risk of asthma attacks; high blood pressure may proceed a myocardial infarction. One may use PROC GENMOD available in SAS for the event history analysis.

Further Readings:
Brown H., and R. Prescott, *Applied Mixed Models in Medicine*, Wiley, 1999.

Factor Analysis

Factor Analysis is a technique for data reduction that is, explaining the variation in a collection of continuous variables by a smaller number of underlying dimensions (called factors). Common factor analysis can also be used to form index numbers or factor scores by using correlation or covariance matrix. The main problem with factor analysis concept is that it is very subjective in its interpretation of the results.

Further Reading:
Reyment R., and K. Joreskog, *Applied Factor Analysis in the Natural Science*, Cambridge University Press, 1996. It covers multivariate analysis and applications to environmental fields such as chemistry, paleoecology, sedimentology, geology and marine ecology.
Tabachick B., and L. Fidell, *Using Multivariate Statistics*, Harper Collins, New York, 1996.

Kinds of Lies: Lies, Damned Lies and Statistics

"There are three kinds of lies -- lies, damned lies, and statistics." quoted in Mark Twain's autobiography.

It is already an accepted fact that "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." However it often happens that people manipulating statistics in their own advantage or in advantage of their boss or friend.

The following are some examples as how statistics could be misused in advertising, which can be described as the science of arresting human unintelligence long enough to get money from it. The founder of Revlon says "In factory we make cosmetics; in the store we sell hope."

In most cases, the deception of advertising is achieved by omission:

1. The Incredible Expansion Toyota: "How can it be that an automobile that's a mere nine inches longer on the outside give you over two feet more room on the inside? May be it's the new math!" Toyota Camry Ad.

   Where is the fallacy in this statement? Taking volume as length! For example: 3x6x4=72 feet (cubic), 3x6x4.75=85.5 feet (cubic). It could be even more than 2 feet!

2. Pepsi Cola Ad.: " In recent side-by-side blind taste tests, nationwide, more people preferred Pepsi over Coca-Cola".

   The questions are, Was it just some of taste tests, what was the sample size? It does not say "In all recent…"

3. Correlation? Consortium of Electric Companies Ad. "96% of streets in the US are under-lit and, moreover, 88% of crimes take place on under-lit streets".
4. Dependent or Independent Events? "If the probability of someone carrying a bomb on a plane is.001, then the chance of two people carrying a bomb is.000001. Therefore, I should start carrying a bomb on every flight."
5. Paperboard Packaging Council's concerns: "University studies show paper milk cartons give you more vitamins to the gallon."

   How was the design of experiment? The council sponsored the research! Paperboard sales is declining!

6. All the vitamins or just one? "You'd have to eat four bowls of Raisin Bran to get the vitamin nutrition in one bowl of Total".
7. Six Times as Safe: "Last year 35 people drowned in boating accidents. Only 5 were wearing life jackets. The rest were not. Always wear life jacket when boating".

   What percentage of boaters wears life jackets? Is it a conditional probability.

8. A Tax Accountant Firm Ad.: "One of our officers would accompany you in the case of Audit".

   This sounds like a unique selling proposition, but it conceals the fact that the statement is a US Law.

9. Dunkin Donuts Ad.: "Free 3 muffins when you buy three at the regular 1/2 dozen price."

   There have been many other usual misuses of statistics: dishonest and/or ignorant survey methods, loaded survey questions, graphs and picto-grams that suppress that which is not in the "proof program," and survey respondents who are the autos select because they have an axe to grind about the issue; very interesting stuff, and, of course, those amplifying that which the data really minimizes.

   Further Readings:

Adams W., *Slippery Math in Public Affairs: Price Tag and Defense*, Dekker. 2002. Examines flawed usage of math in public affairs through actual cases of how mathematical data and conclusions can be distorted and misrepresented to influence public opinion. Highlights how slippery numbers and questionable mathematical conlusions emerge and what can be done to safeguard against them.
Dewdney A., *200% of Nothing*, John Wiley, 1993. Based on his articles about math abuse in Scientific American, Dewdney lists the many ways we are manipulated with fancy mathematical footwork and faulty thinking in print ads, the news, company reports and product labels. He shows how to detect the full range of math abuses and defend against them.
Good Ph., and J. Hardin, *Common Errors in Statistics*, Wiley, 2003.
Schindley W., *The Informed Citizen: Argument and Analysis for Today*, Harcourt Brace, 1996. This rhetoric/reader explores the study and practice of writing argumentative prose. The focus is on exploring current issues in communities, from the classroom to cyberspace. The "interacting in communities" theme and the high-interest readings engage students, while helping them develop informed opinions, effective arguments, and polished writing.
Spirer H., L. Spirer, and A. Jaffe, *Misused Statistics*, Dekker, 1998. Illustrating misused statistics with well-documented, real-world examples drawn from a wide range of areas, public policy, and business and economics.

Entropy Measure

Inequality coefficients used in business, economy, and information processing are analyzed in order to shed light on economic disparity world-wide. Variability of a categorical data is measured by the Shannon-entropy function:

E= - S pi ln(pi)

where, sum is over all the categories and pi is the relative frequency of the ith category. It represents a quantitative measure of uncertainty associated with p. It is interesting to note that this quantity is maximized when all pi's, are equal.

For a rXc contingency table it is E= S S pij ln(pij) - S(S pij) ln(S(pij) - S(S pij) ln(S(pij)

The sums are over all i and j, and j and i's.

Another measure is the Kullback-Liebler distance (related to information theory):

S((Pi - Qi)*log(Pi/Qi)) =
S(Pi*log(Pi/Qi))+ S(Qi*log(Qi/Pi))

or the variation distance

S(| Pi - Qi |)/2

where Pi and Qi are the probabilities for the i-th category for the two populations.

Further Reading:
Kesavan H., and J. Kapur, *Entropy Optimization Principles with Applications*, Academic Press, New York, 1992.

Warranties: Statistical Planning and Analysis

In today global market place, warranty has become an increasingly important component of a product package and most consumer and industrial products are sold with a warranty. The warranty serves many purposes. It provides protection for both buyer and manufacturer. For a manufacturer, a warranty also serves to communicate information about product quality, and, as such, may be used as a very effective marketing tool.

Warranty decisions involve both technical and commercial considerations. Because of the possible financial consequences of these decisions, effective warranty management is critical for the financial success of a manufacturing firm. This requires that management at all levels be aware of the concept, role, uses and cost and design implications of warranty. The aim is to understand the concept of warranty and its uses; warranty policy alternatives; the consumer/manufacturer perspectives with regards warranties; the commercial/technical aspects of warranty and their interaction; strategic warranty management; methods for warranty cost prediction; warranty administration.

Further Reading:
Brennan J., *Warranties: Planning, Analysis, and Implementation*, McGraw Hill, 1994.

Tests for Normality

The standard test for normality is the Kolmogrov-Smirinov-Lilliefors statistic. A histogram and normal probability plot will also help you distinguish between a systematic departure from normality when it shows up as a curve.

Kolmogrov-Smirinov-Lilliefors Test: This test is a special case of the Kolmogorov-Smirnov goodness-of-fit test for normality of population's distribution. In applying the Lilliefors test a comparison is made between the standard normal cumulative distribution function, and a sample cumulative distribution function with standardized random variable. If there is a close agreement between the two cumulative distributions, the hypothesis that the sample was drawn from population with a normal distribution function is supported. If, however, there is a discrepancy between the two cumulative distribution functions too great to be attributed to chance alone, then the hypothesis is rejected.

The difference between the two cumulative distribution functions is measured by the statistic D, which is the greatest vertical distance between the two functions.

Another widely used test for normality is the Jarque-Bera statistic, which is based on the values of skewness and kurtosis of sample data. For large n, (say, over 30) under the normality condition the Jarque-Bera statistic:

$$n \{Skewness2 /6+((Kurtosis - 3)2)/24)\}$$
$$n\{ S32 /(6S23)+ [S4 /(S22 -3)]2 /24 \}$$

follows a chi-square distribution with d.f. = 2, where:

$$S2 = S (xi - \bar{x})2 /(n - 1),$$

$$S3 = S (xi - \bar{x})3 /(n - 1), \text{ and}$$

$$S4 =S (xi - \bar{x})4 /(n - 1).$$

The above test is based on both skewness and kurtosis statistics, the following alternative test is using the kurtosis statistic only:

Let

$$C3 = \{Kurtosis - 3(n-1)/(n+1)\}/\{24n(n-2)(n-3)/[(n+1)2(n+3)(n+5)]\}1/2$$

$$C2 = \{6(n2 - 5n+ 2)/[(n+7)(n+9)]\} \{6(n+3)(n+5)/[n(n-2)(n-3)]\}1/2$$

$$C1 = 6+(8/C2)\{2/C2+ (1+ 4/C2)1/2\}$$

Then the statistic:

$$Z = [1-2/9C1 - \{ (1-2/C1)/(1+ C3\{2/(C1 -4)\}1/2\}1/3]/[2/9C1]1/2,$$

follows the standard normal distribution.

As yet another method, one may use statistic:

$$ZF = (n+ 2)1/2\ (F -3)/3.54$$

that has a standard normal density under the null hypothesis. Where

$$F = 13.29\ Ln(s/t)$$

where s is the standard deviation and/t is mean absolute deviation from $\bar{x}$.

You may like using the well known Lilliefors Test for Normality to assess the goodness-of-fit.

Further Readings

Bonett D., and E. Seierb, A test of normality with high uniform power, *Computational Statistics & Data Analysis*, 40, 435-445, 2002.
Chen G., et al, Statistical inference on comparing two distribution functions with a possible crossing point, *Statistics & Probability Letters*, 60, 329-341, 2002.
Gujarati D., *Basic Econometrics*, McGraw Hill, 2002.
Thode T., *Testing for Normality*, Marcel Dekker, Inc., 2001. Contains the major tests for univariate and multivariate normality.

Directional (i.e., circular) Data Analysis

Directional data analysis also called circular data, are data that are measured on a repeating scale, e.g. the compass or clock. They are used in a wide variety of fields – environmental and geo-science, biology and medicine, military analysis, to mention a few. Standard statistical tools are not useful for such data - for example, the "distance" between 340 and 20 angular degrees is more commonly thought of as 40 degrees, as opposed to the 320 degrees a standard calculation would yield. It covers the exploratory and inferential tools to analyze such data using statistical software experience. Its main applications are in Environmental science for analyzing directional data, propagation and homing patterns, vanishing angles, wind direction, industrial researchers and quality engineers, wheel imbalance, designing and assessing curves in roads and rails, military analysts, tracking aircraft direction, direction of homing signals, targeting performance, biologists and medical researchers, circadian rhythm data.

Further Readings

Arsham H., Kuiper's p-value as a measuring tool and decision procedure for the goodness-of-fit test, *Journal of Applied Statistics*, 15(3), 131-135, 1988.

Kindly e-mail me your comments, suggestions, and concerns. Thank you.

Professor Hossein Arsham

This site was launched on 2/18/1994, and its intellectual materials have been thoroughly revised on a yearly basis. The current version is the 9th Edition. All external links are checked once a month.

Back to Dr. Arsham's Home Page

---

translated by ND

Published (Last edited): 15-05-2012 , source: http://home.ubalt.edu/ntsbarsh/Business-stat/stat-data/Topics.htm

E-commerce
Blog Hosting
Unix
Windows

Web Hosting Reviews

Inmotion Hosting
WebHostingHub
iPage
FatCow
HostGator
Arvixe
GreenGeeks
GoDaddy

Independent reviews and ratings of web hosting services by real customers.